



Risk stratification for coronary surgery

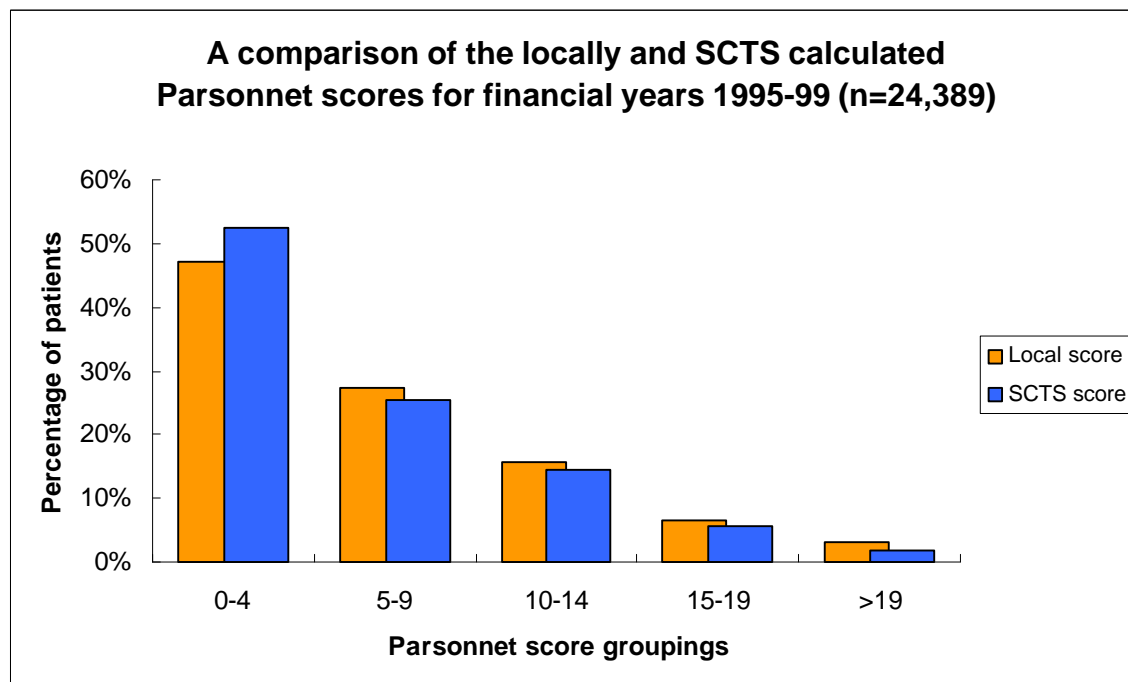
Not all patients are the same. The risk that any one patient will not survive surgery is dependent on a number of different factors, some of which can be quantified, such as age, gender and the existence of co-morbidities. Risk scoring systems attempt to take account of these risk factors and convert them into a numeric risk score. The higher the score, the greater the predicted risk. However, “low risk” is not the same as “no risk”.

Over the years a variety of risk stratification systems have evolved using logistic regression and Bayes modelling techniques^{9,10,11}. These range from simple additive systems^{12,13,14,15} to highly complex statistical algorithms. While none of these systems can accurately predict the outcome for an individual patient some models are better than others in estimating risk for cohorts of patients. This provides the basis for rational and meaningful comparisons of outcomes between groups of patients, institutions and individual surgeons by taking patient related variables and co-morbidity into account.

The Parsonnet score

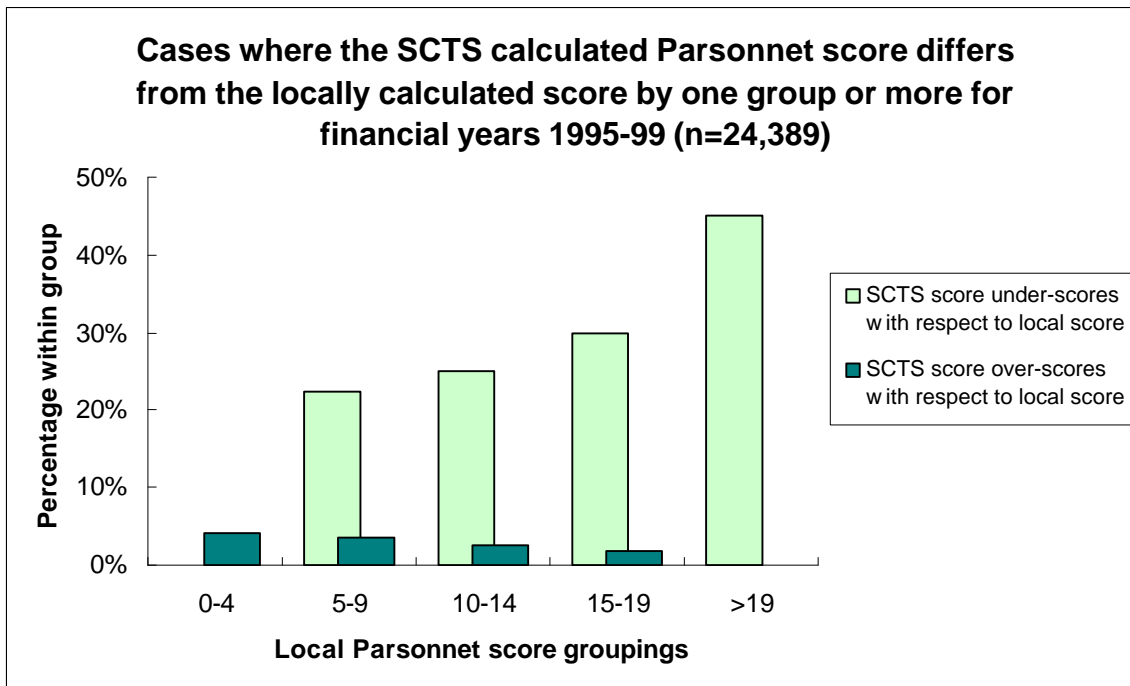
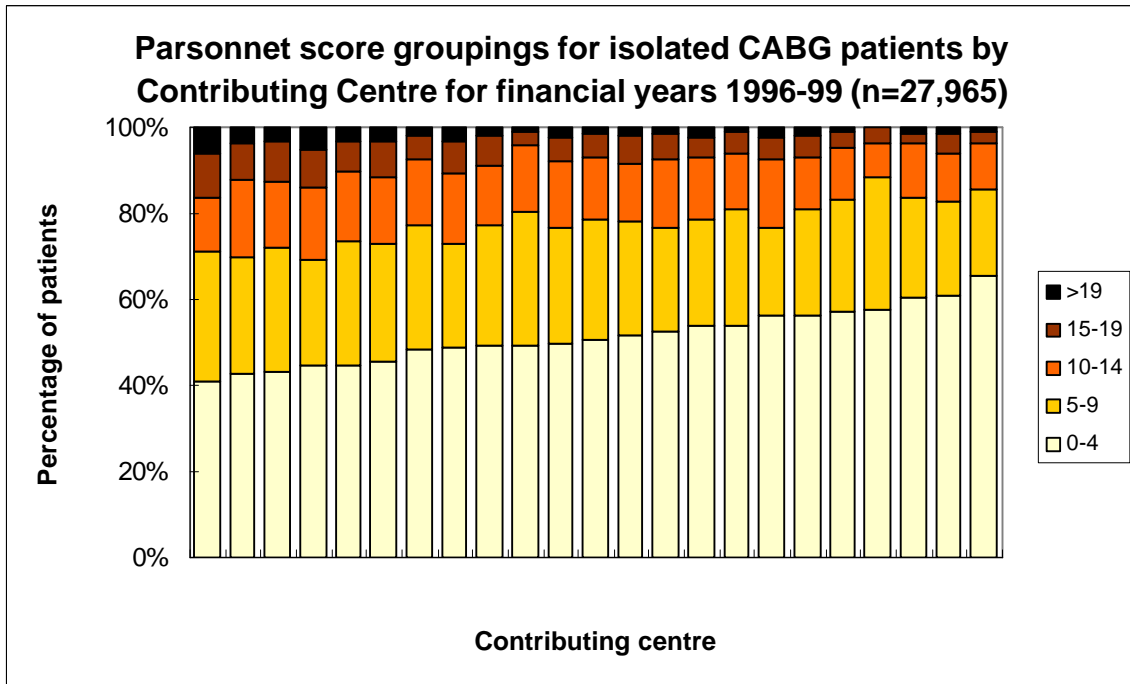
The components that comprise the additive Parsonnet scoring system¹² are detailed in Appendix 2. For those who are unfamiliar with this scoring system, a score for each patient is calculated before the operation and the greater the score, the greater the mortality risk. In the paper published by Parsonnet in 1989, patients in the “low risk” group (Parsonnet score of 0 to 4) had an average mortality of 1%, patients in the “elevated risk” group (Parsonnet score of 5 to 9) had an average mortality of 5%, and so on up the scale. The distribution of patients between the five Parsonnet score groups is usually such that most patients have low scores, and as the risk score increases so the number of patients with that score falls.

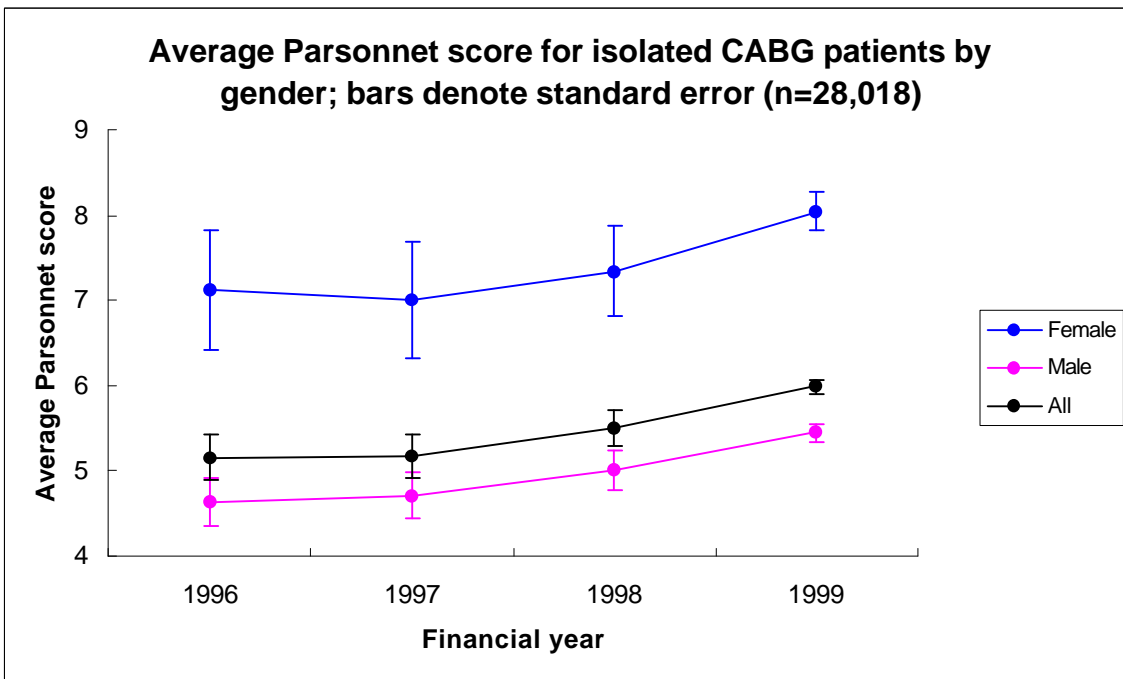
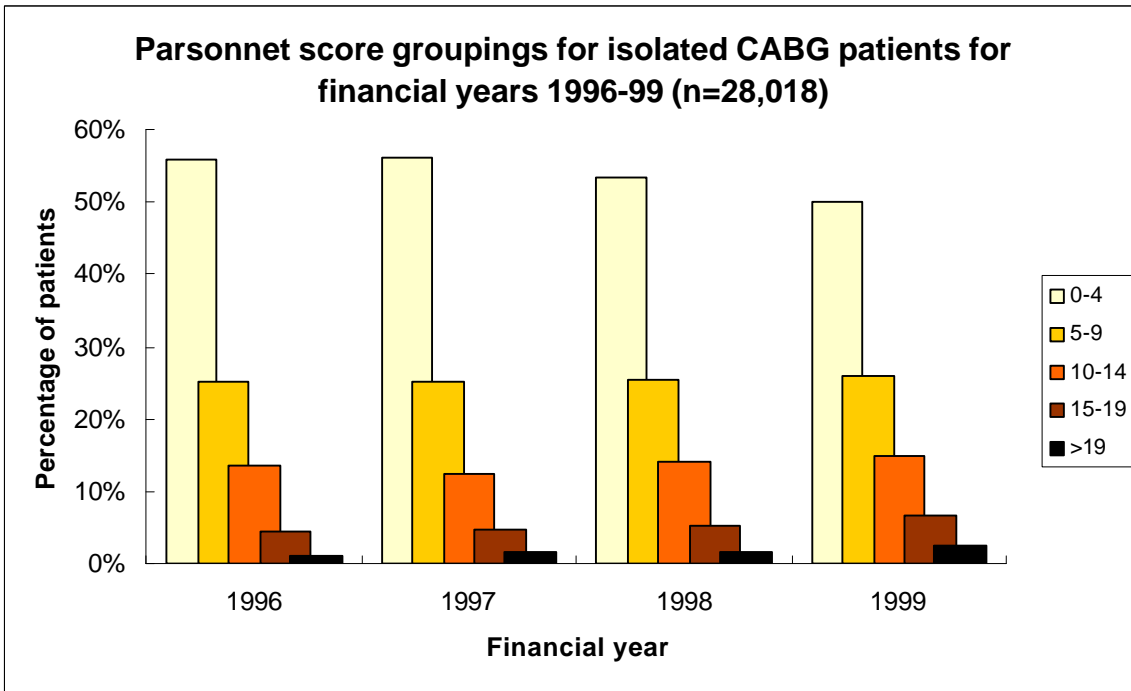
The Parsonnet score can be calculated using all the parameters that Parsonnet described, or by using only the rigidly defined parameters *i.e.*, excluding the subjective “catastrophic states” and “other rare circumstances” (Appendix 2). The second, more stringent approach, which excludes the ill-defined components, was applied to those data entries where all the appropriate data were available to generate the centrally calculated Parsonnet score. This score was then compared to the Parsonnet score recorded by each of the cardiac centres (locally calculated Parsonnet score). The graph and table below demonstrate that the two scores match exactly in terms of the groupings for 83% of patient entries.

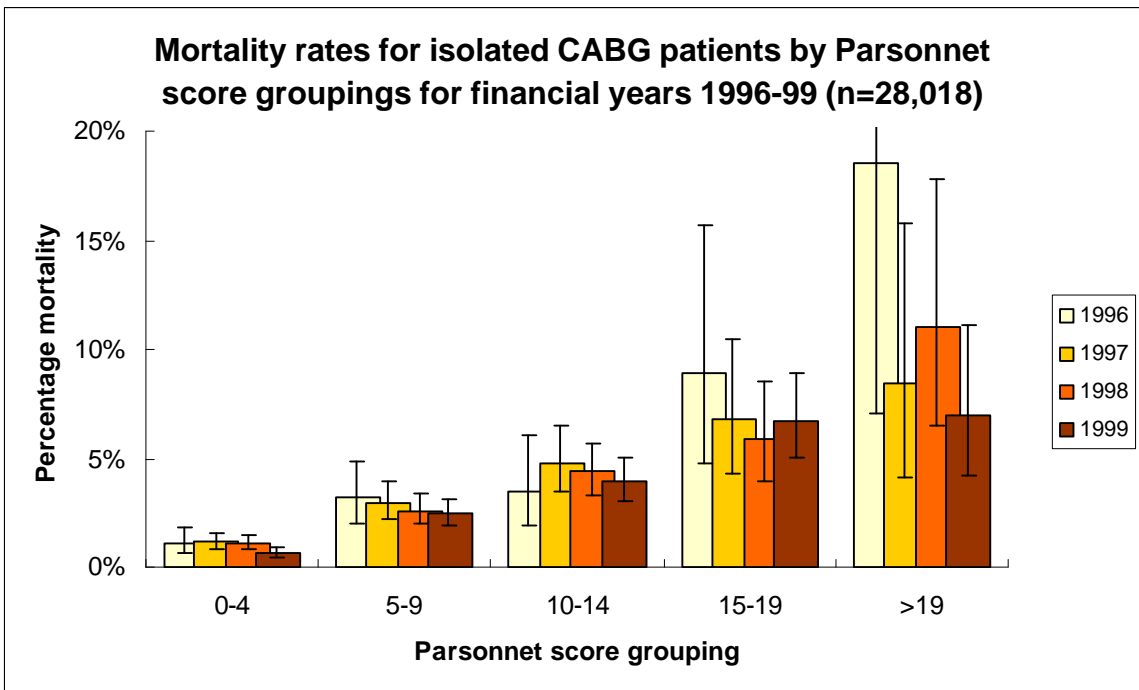
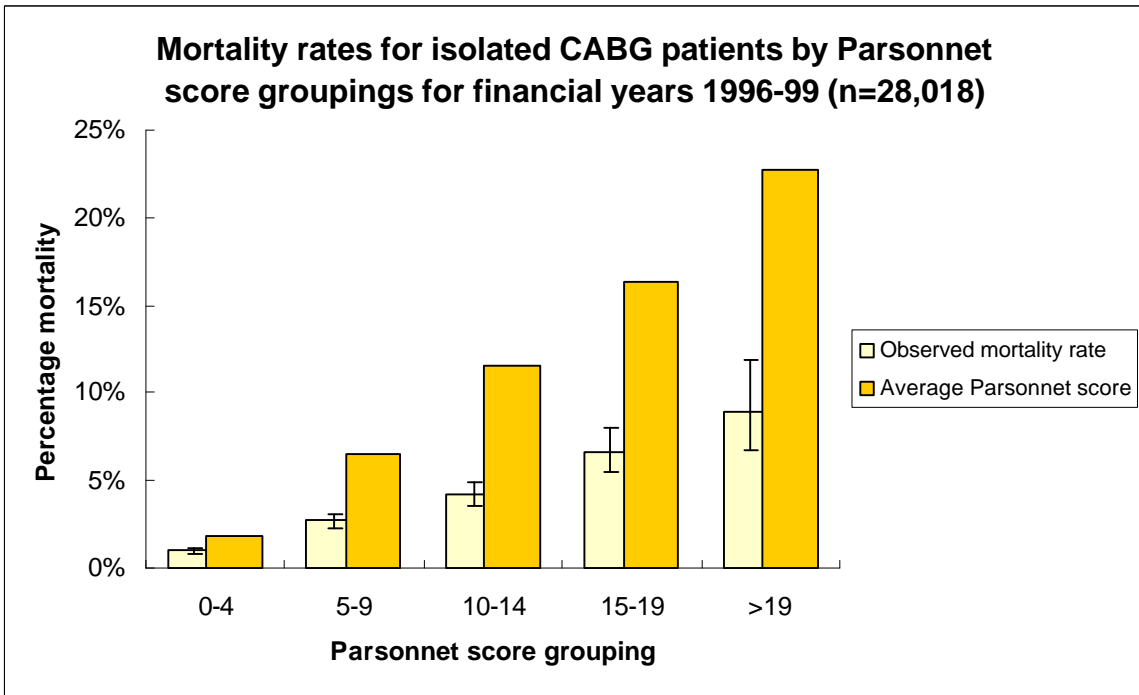




From this point onwards, the Parsonnet score referred to will be the centrally calculated version, so that any comparisons will be made in the knowledge that there has been a standardisation.







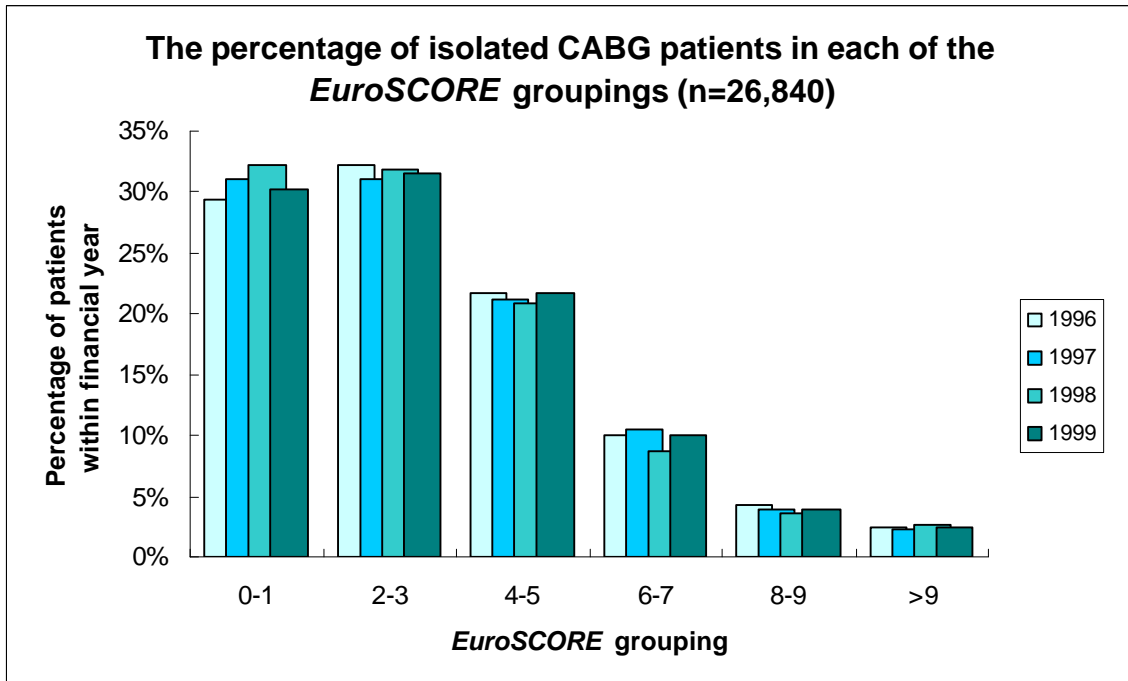


EuroSCORE

The Parsonnet system was devised in North America in the late 1980's. Since then the specialty has moved forward and although most of the risk variables in the Parsonnet system remain pertinent their relative impact on mortality has changed. More recently another system based on a pan-European patient population has been described in order to make the system more applicable to European patients^{16,17}. The principle is much the same, but some of the risk factors and their weightings are different, making allowance for advances in surgical practice and a different patient population.

We have calculated the *EuroSCORE* for a group of coronary artery bypass patients from the current data in the Adult Cardiac Surgical Database (Appendix 3). A total of 16,060 patient entries fulfilled all the criteria required to determine the *EuroSCORE*. Two assumptions were made in the calculation: previous cerebrovascular accidents (CVA) were taken as an indication that there was neurological dysfunction, and the time-frame for a recent heart attack (myocardial infarction) was within the last 30 days rather than within the 90 days stipulated in the *EuroSCORE* calculation.

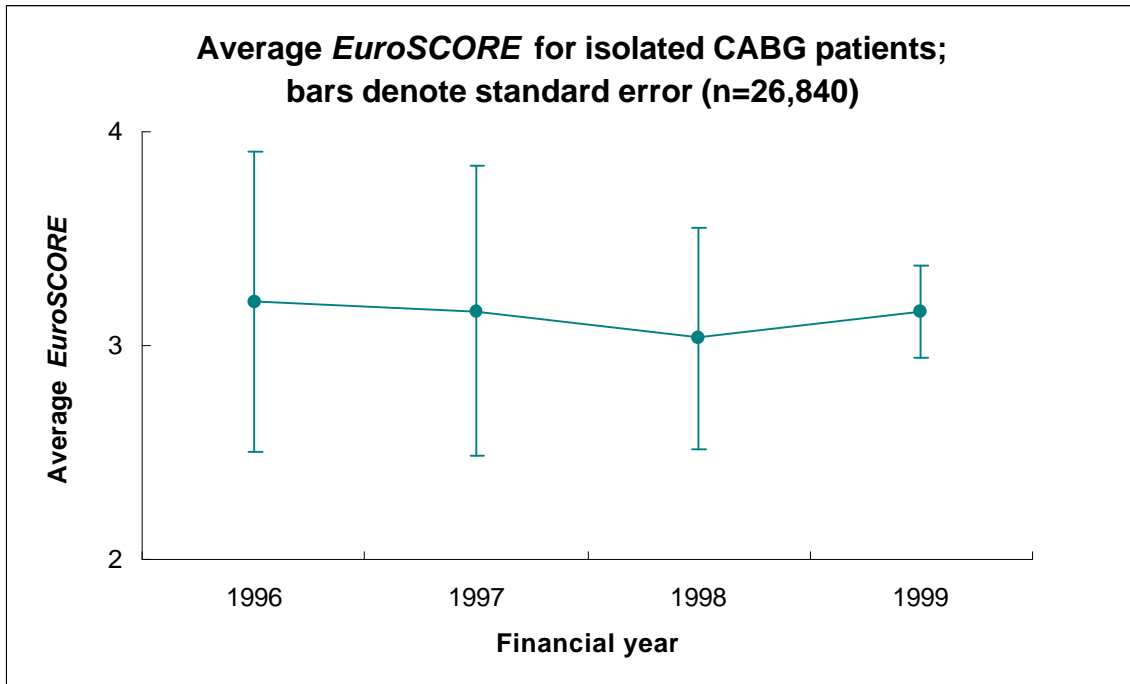
The *EuroSCORE* is a more direct measure of operative mortality than the Parsonnet score. Most patients have a score or between 0 or 3 which approximates to a risk of death in the range 0-3%. This is in keeping with the average mortality rate for contemporary coronary artery bypass surgery.



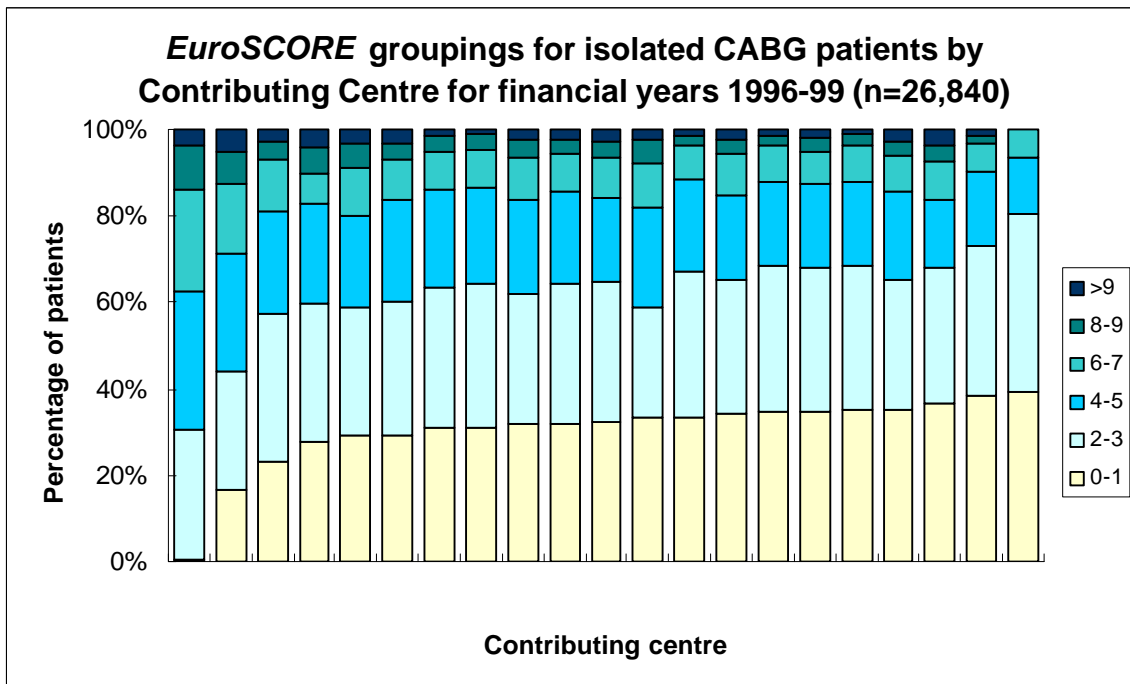
The shape of this distribution is almost identical to that described in the previous National Adult Cardiac Surgical Database report. Almost two thirds of patients have a *EuroSCORE* in the range 0-3, and only 6.4% have a *EuroSCORE* in excess of 7.

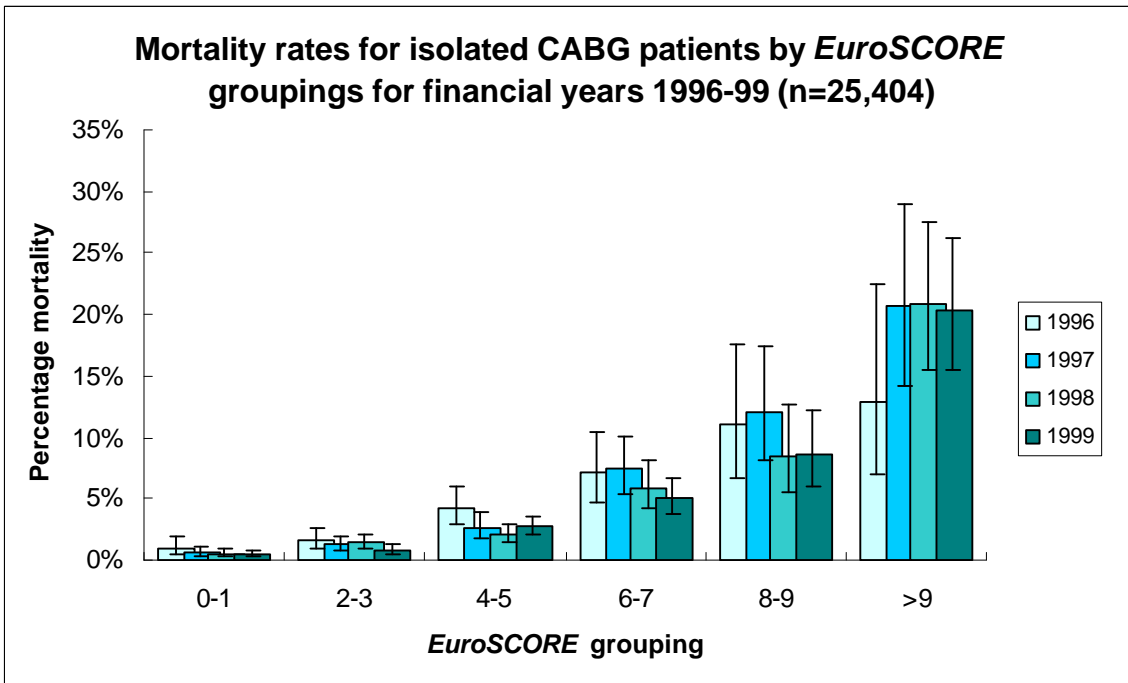
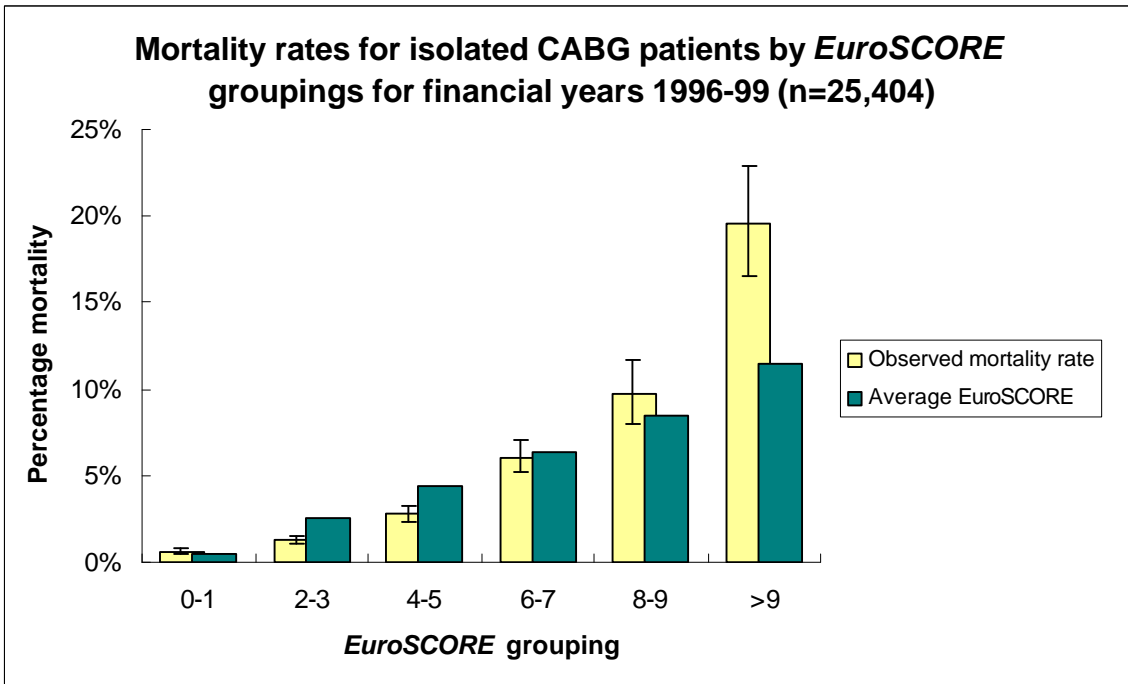


There was no discernable trend in the average *EuroSCORE* over the period encompassed by the financial years 1996 to 1999.



As with the Parsonnet score, there are differences in the distribution of the *EuroSCORE* between centres.







Bayesian analyses

“The sciences ... mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is that it is expected to work.”

John Van Neumann

Risk stratification is a method of delimiting sub-populations within a cohort that have different risks of a particular outcome, based upon severity of illness and co-morbidity. Using such an approach, it is possible to make “fair” comparisons between different institutions or different surgeons. Comparisons of individual institutions’ or individual surgeons’ outcome rates could also be made against agreed standards using this method. The Bayesian approach is one method of risk stratification.

What do Bayes tables do?

The Bayes table approach is a particularly simple way of building a risk stratification system from a database. Based solely on tables relating outcomes to single risk factors, the probability of an adverse outcome can be estimated for a patient with any combination of risk factors.

The method is based on the repeated use of Bayes theorem, which is a basic formula in probability theory, first discovered by the Rev. Thomas Bayes, a non-conformist minister from Tunbridge Wells, which was published posthumously in 1763 (Thomas Bayes having died in 1761; his grave lies in Bunhill Fields in London, only yards from the Royal Statistical Society). Bayes theorem tells us how the probability of an event should be revised when additional relevant evidence is obtained.

The following is an example of the way in which this may be applied to clinical data. Suppose the event of interest is the outcome of post-operative death, and we have information on both post-operative status and age. The data might be shown as in the table below:

Table of fictitious, example data

		<i>Post-operative status</i>		
		<i>S: Survivors</i>	<i>D: Deaths</i>	<i>Total</i>
<i>Age grouping</i>	<i>a: 70-79 years</i>	90	30	120
	<i>not a: other ages</i>	810	70	880
	<i>Total</i>	900	100	1,000

The notation used within the table is such that *S* represents survival, *D* represents death, *a* represents patients in the age grouping 70-79 years old and *not a* represents all other patients.

A probability of an event is calculated as the number of events in a group (post-operative deaths) divided by the total number in that same group. For example, the probability of death, $p(D)$, for the entire group is $100 / (100 + 900) = 0.10$. Such rates can also be expressed in percentage terms, and the percentage rate is obtained by multiplying the probability by a factor of 100. In this example, the percentage mortality rate is $[100 / (900 + 100)] \times 100 = 0.10 \times 100 = 10\%$. The probability of survival, $p(S)$, is $900 / (100 + 900) = 0.90$, which may also be expressed as 90%.

A probability of 1.00 (100%) means that the event always occurs, whereas a probability of 0.00 (0%) means that the event never occurs. Since either death or survival must occur, the sum of the probability of survival and the probability of death must be equal to 1. This idea may be expressed as the formula:

$$p(D) + p(S) = 1$$

This implies that the probability of survival can be calculated as:

$$p(S) = 1 - p(D)$$

The odds on death is defined as the probability of death divided by the probability of survival

$$\text{Odds } (D) = p(D) / p(S) = p(D) / [1 - p(D)]$$



In the above example, the odds on death for the whole group is $p(D) / [1 - p(D)] = 0.1 / [1 - 0.10] = 1 / 9$ or, in betting parlance, 9 to 1 against death. An assessment of these odds based on no patient specific data is known as the *prior odds*.

The odds on D may be converted to a probability using the following formula:

$$p(D) = \text{Odds}(D) / [1 + \text{Odds}(D)]$$

Suppose that we now wish to take into account the age of the patient, which falls in the range 70-79 years old (designated a). The probability of death in this group is denoted $p(D | a) = 30 / (90 + 30) = 30 / 120 = 0.25$ or 25%. This is known as the *posterior probability*. The odds on death in this group is denoted $p(D | a) / p(S | a) = 0.25 / 0.75 = 1 / 3 = 0.33$ or 3 to 1 against death. This is known as the *posterior odds*.

Bayes theorem is the formula that provides the link between the prior and posterior odds:

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}$$

where the likelihood ratio expresses how much more likely it is that a patient with such an age should fall amongst those who die rather than those who survive. From the above, the likelihood ratio = $p(a | D) / p(a | S) = [(30 / 100) / (90 / 900)] = 0.30 / 0.10 = 3$, *i.e.*, this age group is 3 times more common in those who die than those who survive. Thus, Bayes theorem suggests that the posterior odds = $3 \times \text{prior odds} = 3 \times (0.10 / 0.90) = 3 \times 0.11 = 0.33$. Using the conversion formula described above, this corresponds to a probability of 0.25 or 25%, which matches the probability obtained by direct examination of the data in the table.

The Bayes tables, on pages 135 - 138, give detailed information on Bayesian risk models. The Bayes tables in this report have a column labelled Odds Ratio. The data in this column for the risk factor group OVERALL correspond to the prior odds, and the data for the remaining risk factor groups are the likelihood ratios. The column labelled weight is calculated as $10 \times \ln(\text{likelihood ratio})$.

There is an extended, practical example of the way in which surgeons can calculate the risk for an individual patient who has a particular set of risk factors on page 128 of this report.

Items of evidence in a Bayes model might include many different risk factors that affect the outcome, such as age, left ventricular ejection fraction, urgency and so on. The calculation of risk for a specific patient is updated each time a new item of evidence is added. This simple procedure is, however, making the crucial assumption that each item of evidence is contributing independent information concerning the chances of the outcome; technically, we are assuming that the items of evidence are conditionally independent given the true outcome. This assumption is most likely to be appropriate if careful clinical sense has been used in selecting the predictive factors that do not convey similar evidence. Failure to select predictive factors in this way may result in predictions that are over-confident.

It is important to note that the score is calculated irrespective of omissions in the data. It is possible to calculate a Bayes score whether all, some or none of the risk factors are known. The Bayes score is adjusted for each item of evidence, and will more accurately reflect the true risk if all the relevant data are known.



Two Bayesian risk models for isolated coronary artery bypass surgery

Two Bayesian models were created to predict in-hospital mortality from the data on isolated coronary artery bypass surgery: a simple, 5-factor Bayes model and a complex, 9-factor Bayes model. Simple risk models can produce acceptable results, but more complex risk models may give better results. These two risk models were generated in an attempt to test this hypothesis.

The aim was to generate risk models that accurately discriminated between patients who died following surgery and patients who survived; this was tested using Receiver Operating Characteristic curve analysis (see below). The risk score should also provide an accurate estimation of the individual patient's risk; this was tested using calibration curve analyses (see below). Each Bayes model was initially trained on the isolated coronary artery bypass surgery data from the financial year 1998-1999 (Appendices on pages 135 and 137), and then tested on the isolated coronary artery bypass surgery data from the financial years 1998-1999 and 1999-2000 separately. When results from the ROC curve and calibration plot analyses indicated that the models discriminated well and produced broadly accurate predictions, both were re-calculated on the pooled data from 1998-2000 (Appendices on pages 136 and 138).

The results from both these tests were compared to similar results for the other common, additive risk models: the Parsonnet score and the *EuroSCORE*.

A simple, 5-factor model for isolated CABG procedures

The first model contained only 5 risk factors, in an attempt to provide a reasonably simple predictor of the outcome that all surgeons could use. A minimum of data is needed to calculate the score.

A large number of risk factors were ranked according to their weight-of-evidence. The weight-of-evidence for each risk factor was determined as the sum of the products, for each subdivision of the risk factor, of the normalised proportion of patients in the death column and the normalised weight in the Bayes table. In descending order of weight-of-evidence the risk factors were: ejection fraction, priority, cardiogenic shock, age, congestive cardiac failure, any pre-operative support (determined from the presence from cardiogenic shock, IV inotropes, intra-aortic balloon pump, pre-operative ventilation), angina class, angina symptom status, IV inotropes, left main stem disease, intra-aortic balloon pump, number of previous operations, IV nitrates, dyspnoea grade, pre-operative ventilation, peripheral vascular disease, renal dysfunction, body surface area, previous myocardial infarction, pacemaker, recent failed intervention, extent of coronary disease, gender, pulmonary disease, cerebrovascular disease, diabetes, hypercholesterolaemia, hypertension. The patient's ejection fraction adds the greatest weight-of-evidence, and the presence or absence of hypertension adds the least weight-of evidence.

A series of 5-factor Bayes models were constructed, using various combinations of risk factors, giving preference to those with the greatest weight-of-evidence. The combination of risk factors that gave the best results was: age (weight-of-evidence rank 4), body surface area (weight-of-evidence rank 18), ejection fraction (weight-of-evidence rank 1), priority (weight-of-evidence rank 2) and previous operations (weight-of-evidence rank 12) (Appendices on pages 135 and 136). Calculation of this score requires seven data items: the patient's date of birth, the date of the operation, the patient's height and weight, ejection fraction, the urgency of the operation and whether or not the patient had undergone any cardiac surgery previously.

Taking risk factors on the basis of their weight-of-evidence alone did not give the best results.

A complex, 9-factor model for isolated CABG procedures

The second model contained 9 risk factors: age (weight-of-evidence rank 4), body surface area (weight-of-evidence rank 18), diabetes (weight-of-evidence rank 26), hypertension (weight-of-evidence rank 28), left main stem disease (weight-of-evidence rank 10), ejection fraction (weight-of-evidence rank 1), priority (weight-of-evidence rank 2), renal system (weight-of-evidence rank 17) and previous operations (weight-of-evidence rank 12) (Appendices on pages 137 and 138). Eleven data items are required to calculate this score: the patient's date of birth, the date of the operation, the patient's height and weight, the presence or absence of diabetes, hypertension, left main stem disease and renal disease, the patient's left ventricular ejection fraction, the urgency of the operation and whether or not the patient had undergone any cardiac surgery previously.

Again, taking risk factors simply on the basis of their weight-of-evidence did not give the best results.



Testing a risk scoring system

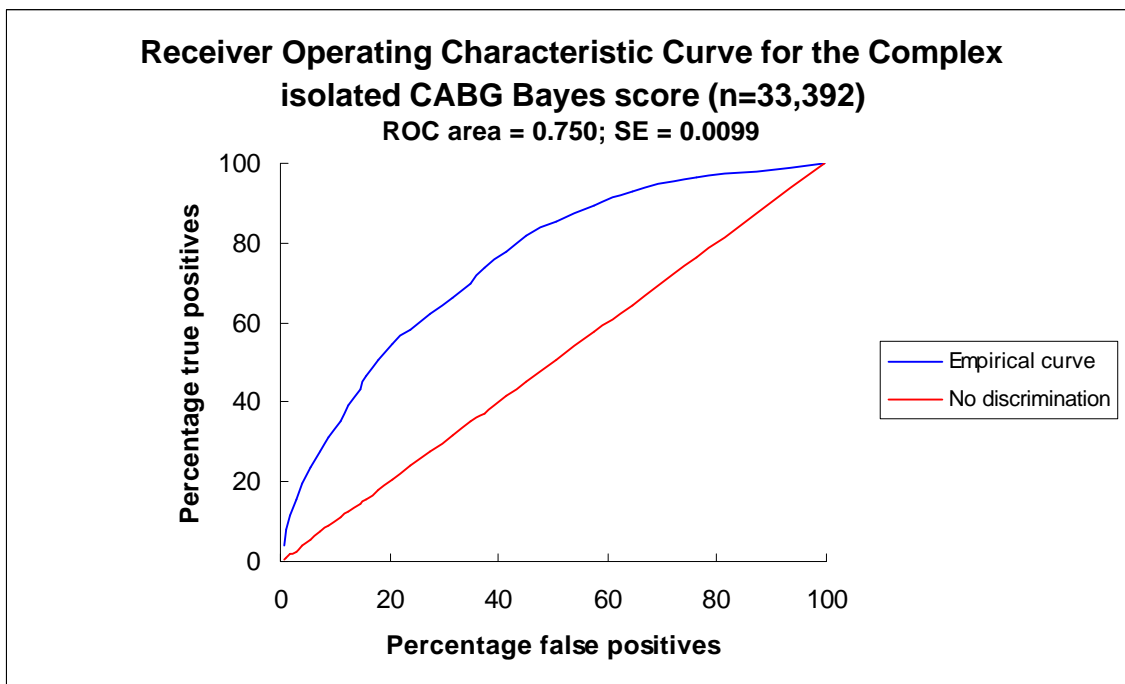
Although there are a number of risk stratification models available, the Parsonnet and Bayes scores are used most frequently. The *EuroSCORE* is a new model that is under assessment but seems to accurately predict outcomes in contemporary UK practice. Which is the most accurate system? Statisticians have a number of methods for measuring the predictive accuracy of such scoring systems. The Receiver Operating Characteristic (ROC) curve^{18,15} and calibration plots provide useful graphical representation of predictive accuracy.

Receiver Operating Characteristic curve

Hanley and McNeil stated (with our insertions in square brackets) that the area under the ROC curve:

"... represents the probability that a randomly chosen diseased subject [or in this case a deceased patient] is (correctly) rated or ranked with greater suspicion than a randomly chosen non-diseased subject [or in this case a non-deceased patient]."

A simplistic reworking of this statement would be that the area represents the probability that the risk predictor [in this report the Parsonnet score, the *EuroSCORE* or the Bayes score accurately discriminates between patients who die during surgery and patients who survive surgery. An area of 0.50 indicates that there is no discrimination *i.e.*, individuals in survivor-deceased patient pairings are allocated to the correct group by the risk predictor according to chance. An area of 1.00 would indicate that discrimination was perfect, and any intermediate value is a quantitative measure of the ability of the risk predictor to distinguish between survivors and non-survivors. Obviously, the closer the value is to 0.50 the less accurate the discrimination, and the closer to 1.00 the better the discrimination. By way of an example, the ROC curve for the complex CABG Bayes score is shown below.



Calibration plots

Risk scores must also provide an estimate of the risk for both individual patients and groups of patients. One way to test this component of a risk score is to plot the observed number of events against the predicted number of events, and this is termed a calibration plot. To simplify the procedure, the data can be split into groups, according to risk, and the observed and predicted outcome rates plotted side-by-side. If the model accurately predicts the outcome, the two should match closely. Calibration plots for the Parsonnet score and the *EuroSCORE* have been shown previously.

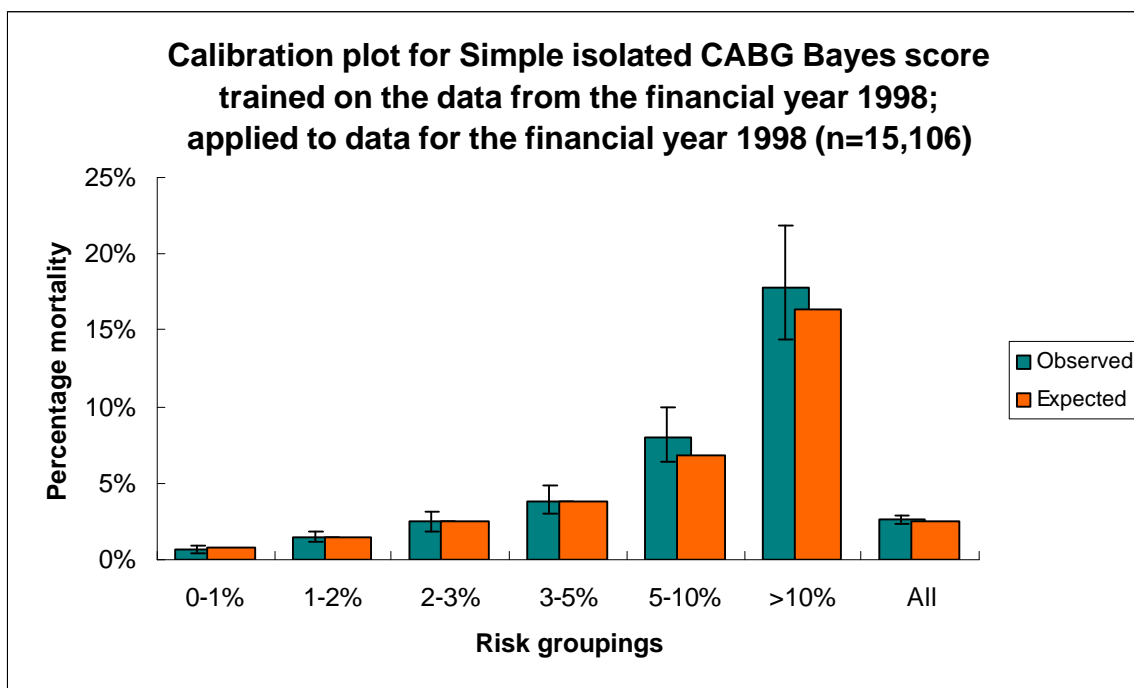
**Testing the Bayes models***The simple Bayes model*

The Receiver Operating Characteristic curve analysis showed that the simple Bayes model discriminated reasonably well. The discrimination was good for the data on which the model was initially trained and for the following year's data. This tended to suggest not only that the model discriminated well, but also that the model was robust. The discrimination was retained when the model was re-calculated on the combined 1998 and 1999 data.

ROC curve areas and standard errors for the simple, 5-factor Bayes model

		Financial year		
		1998	1999	1998 & 1999
ROC areas	Trained on 1998 data	0.754	0.734	N.D. ⁱ
	Re-trained on 1999 data	0.754	0.736	0.744
ROC standard errors	Trained on 1998 data	0.0143	0.0136	N.D. ⁱ
	Re-trained on 1999 data	0.0143	0.0136	0.0099

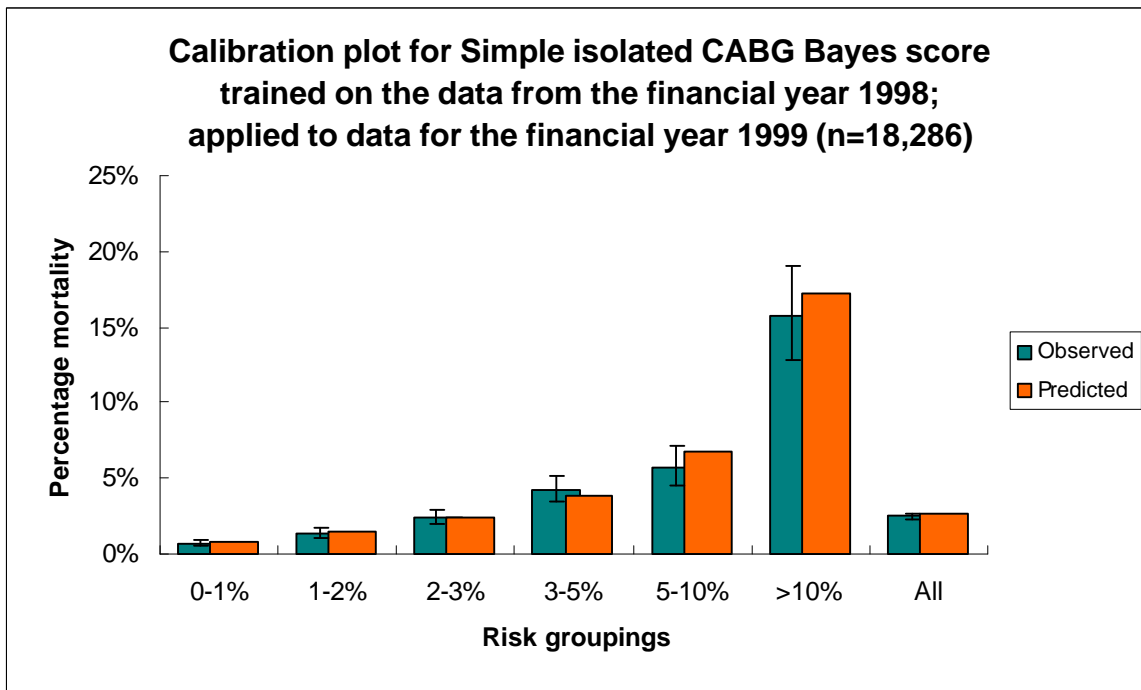
The calibration curve on the 1998 data, the data used to generate the model, showed that the agreement between the observed mortality rate and the predicted mortality rate was good.



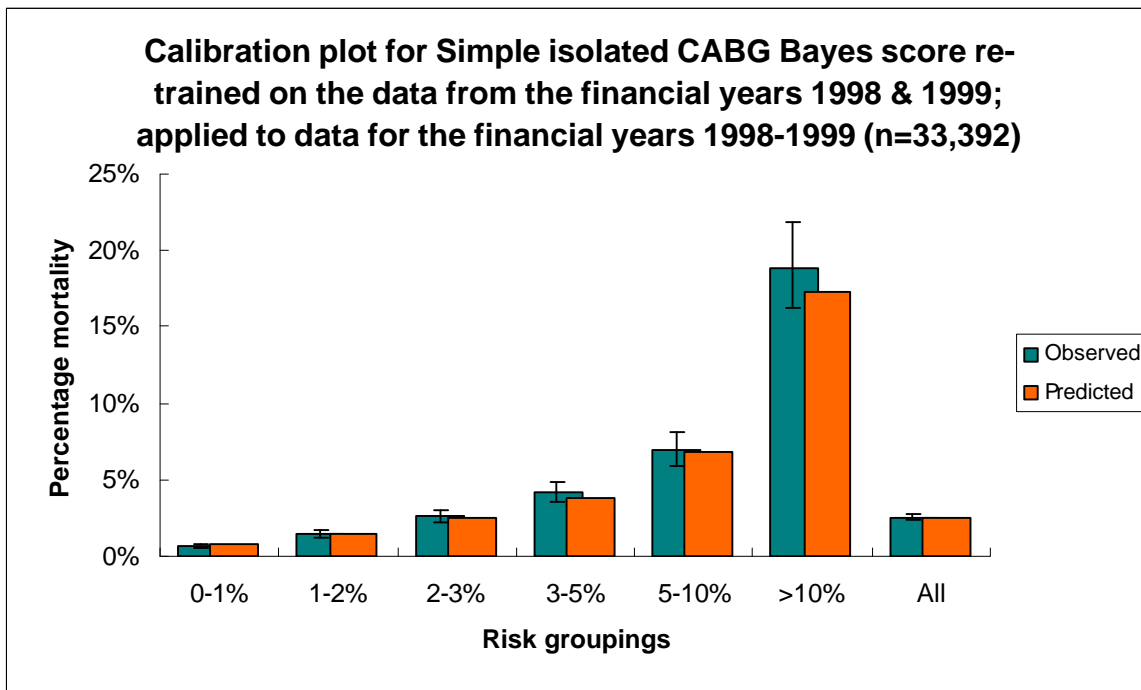
ⁱ N.D. – not determined



The score also calibrated well on the next year's data:



After the model had been recalculated, the calibration plot for the two financial years 1998-1999 and 1999-2000 showed that the prediction improved.





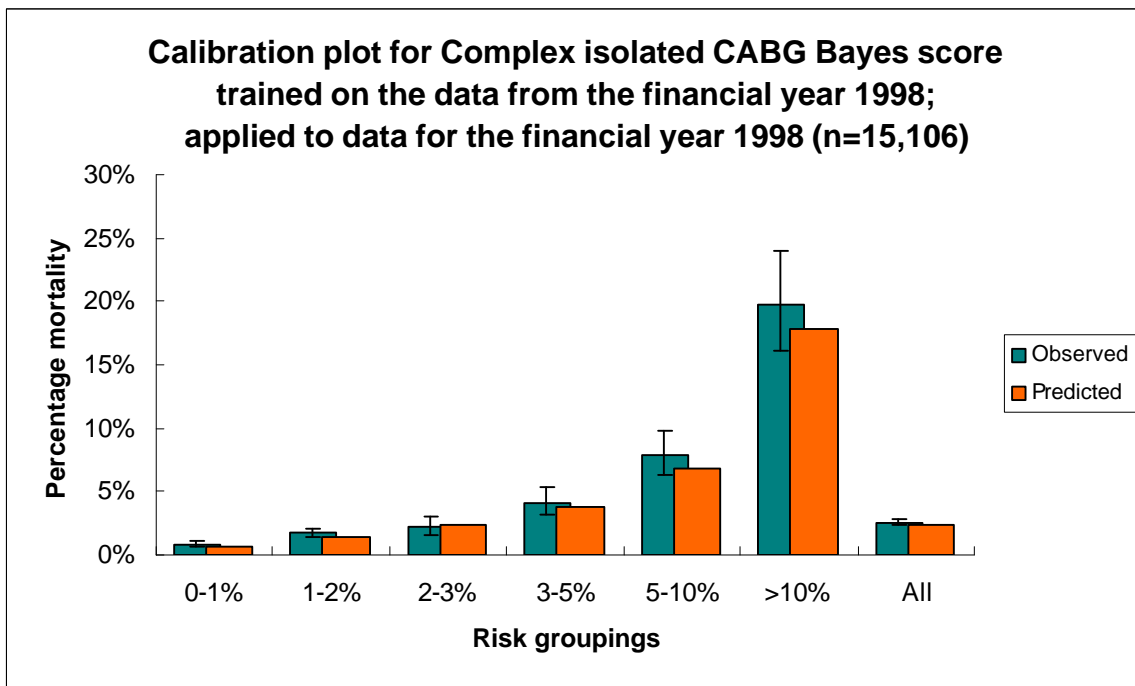
The complex Bayes model

The results for the complex Bayes score were very similar to those for the simple model. The Receiver Operating Characteristic curve analysis showed that the complex Bayes model discriminated reasonably well. The discrimination was good for the data on which the model was initially trained and for the following year's data. This suggested not only that the model discriminated well, but also that the model was robust. The discrimination was retained when the model was re-calculated on the combined 1998 and 1999 data.

ROC curve areas and standard errors for the complex, 9-factor Bayes model

		Financial year		
		1998	1999	1998 & 1999
ROC areas	Trained on 1998 data	0.758	0.741	N.D. ⁱ
	Re-trained on 1999 data	0.757	0.744	0.750
ROC standard errors	Trained on 1998 data	0.0143	0.0136	N.D. ⁱ
	Re-trained on 1999 data	0.0144	0.0136	0.099

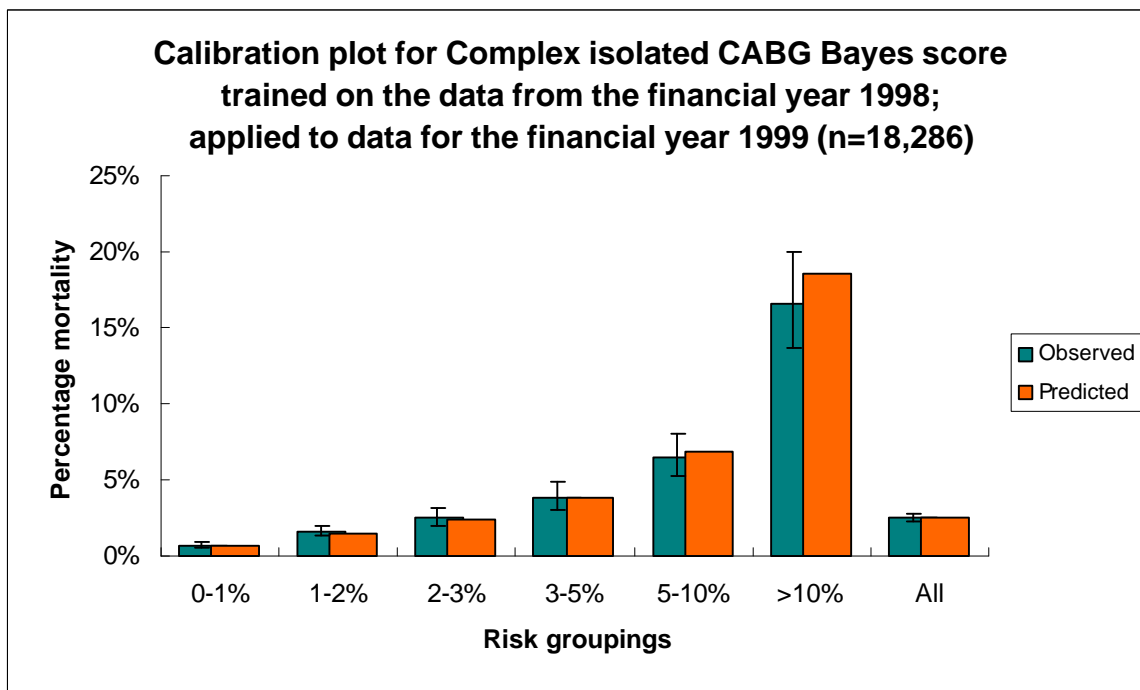
The calibration curve on the 1998 data, the data used to generate the model, showed that the agreement between the observed mortality rate and the predicted mortality rate was good.



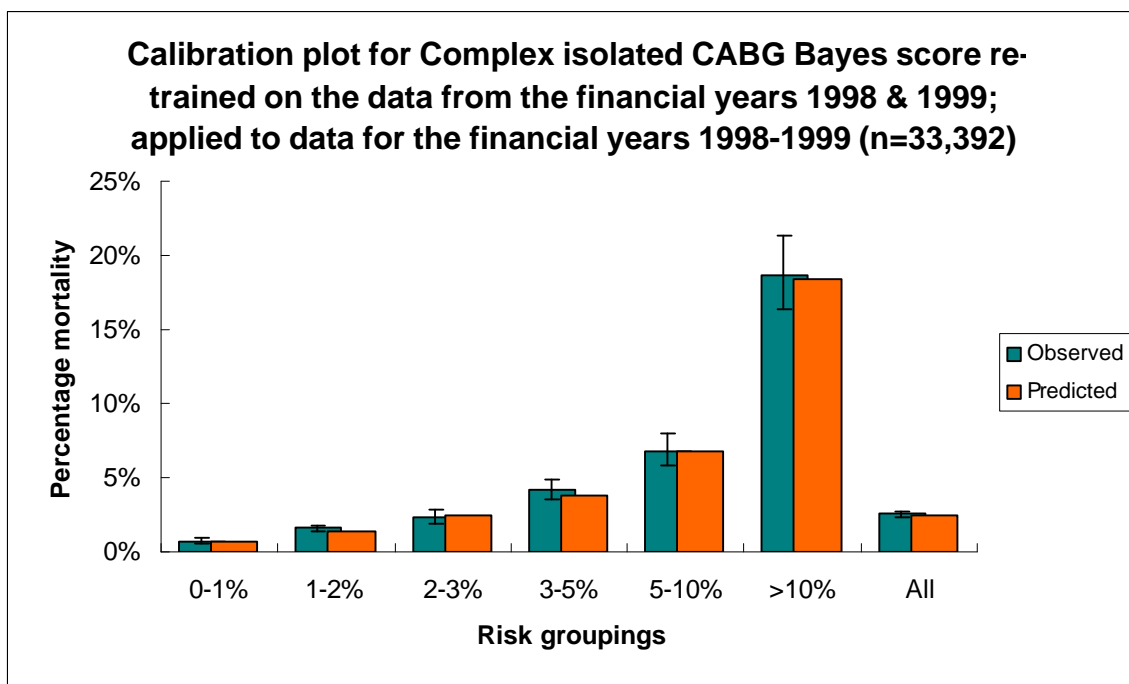
ⁱ N.D. – not determined



As with the simple Bayes score, the score also calibrated well on the next year's data:



After this model had been recalculated on the 1998 and 1999 financial year data, the calibration plot for the two financial years 1998-1999 and 1999-2000 showed that the prediction improved.



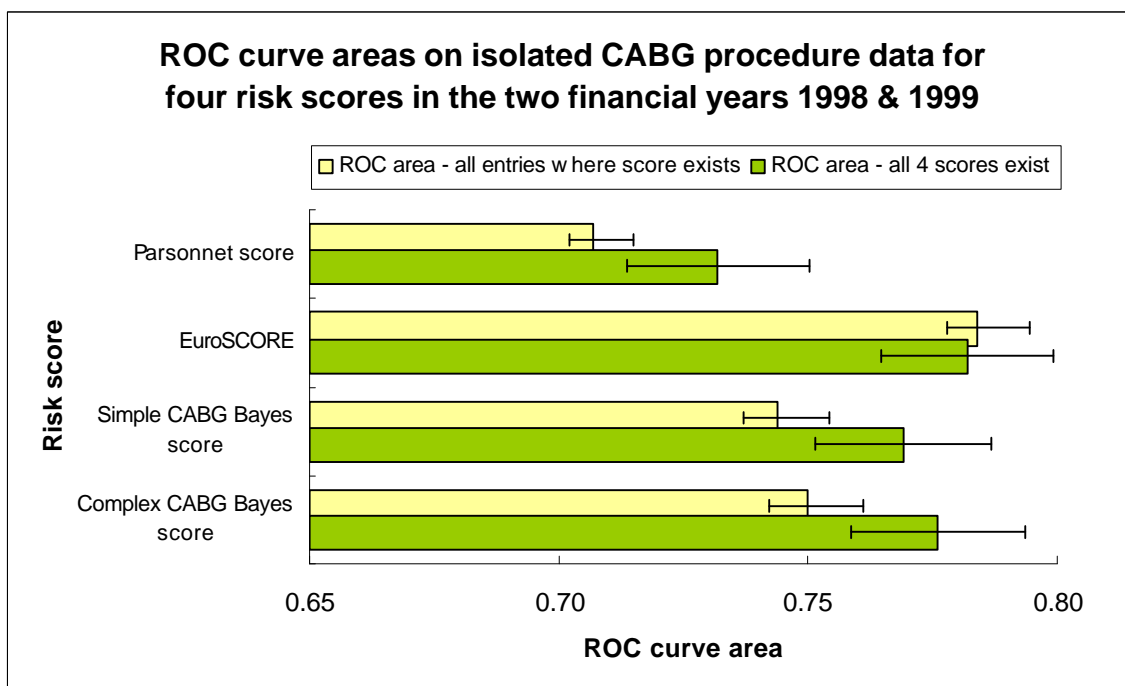


A comparison of the Bayes scores with other risk models

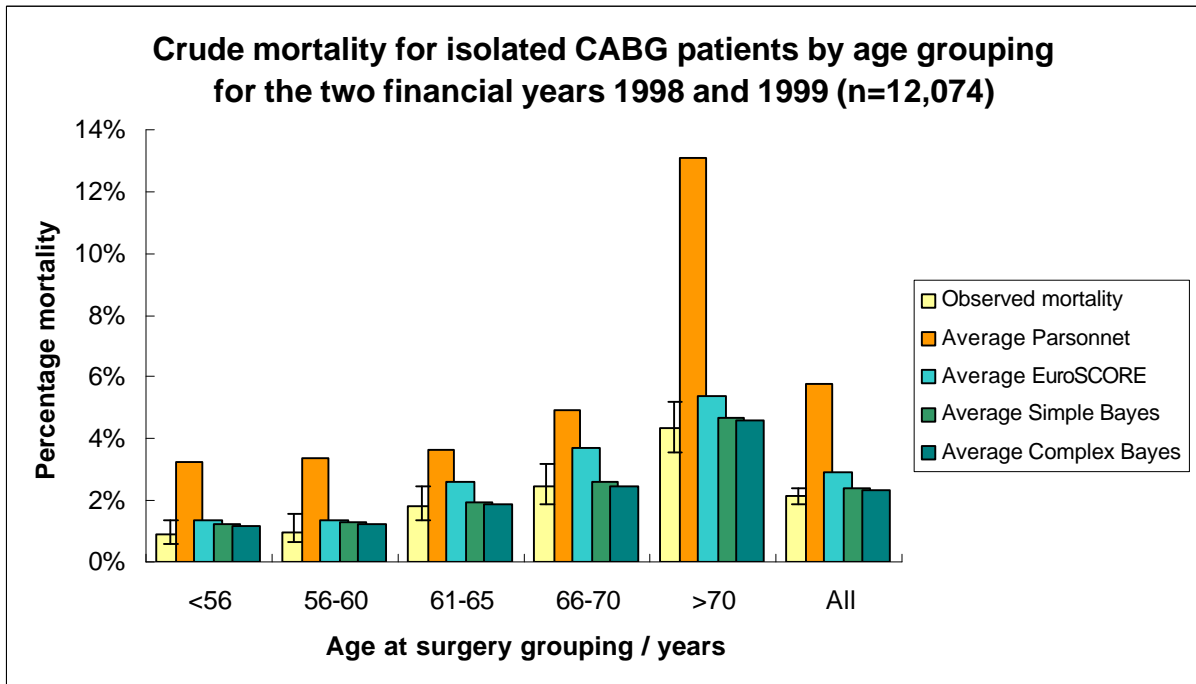
ROC analyses, summarised in the table and chart below, show that the Parsonnet score provides reasonable discrimination. The two Bayes scores provide much better discrimination than the Parsonnet score, but it is the *EuroSCORE* that provides the best discrimination.

ROC curve areas, standard errors and number of entries used in the analysis for four risk scores designed to predict mortality

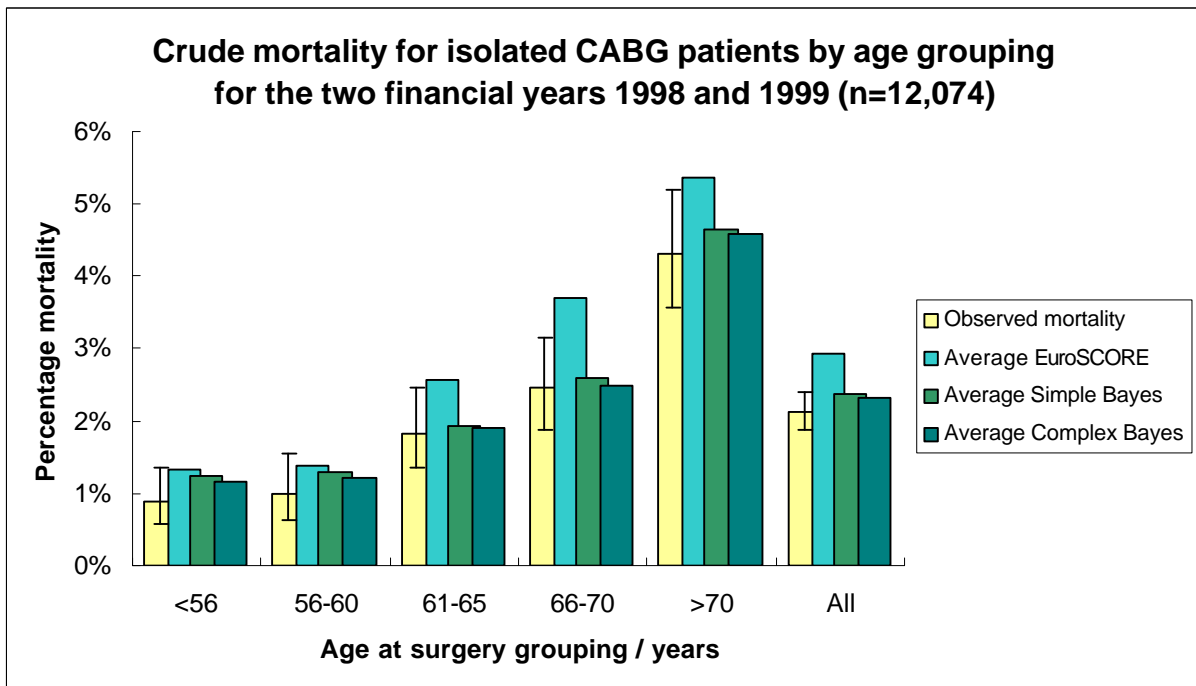
			Risk score			
			Parsonnet score	Euro SCORE	Simple CABG Bayes score	Complex CABG Bayes score
Data used in analyses	All entries where the score exists	ROC area	0.707	0.784	0.744	0.750
		Standard error	0.0145	0.0138	0.0099	0.0099
		Number of entries	17,343	16,907	33,392	33,392
	Entries where all four scores exist	ROC area	0.732	0.782	0.769	0.776
		Standard error	0.0182	0.0173	0.0176	0.0175
		Number of entries	12,074	12,074	12,074	12,074



The second component of the comparison is to determine whether or not the various scores accurately predict mortality. Discrimination is a vital characteristic of a risk score, but if that score is to be used to monitor performance, then it must also reflect true risk. The Parsonnet score, first published in 1989, no longer estimates risk accurately, as seen from the chart overleaf. The *EuroSCORE* represents an improvement on the Parsonnet score in that it provides an approximate estimation of current outcome rates. The Bayes scores, trained on the most recent UK cardiac surgical data available, provide an even more accurate prediction; the complex, 9-factor Bayes score represents a slight improvement on the simpler Bayes model.



When the column showing the mortality prediction according to the Parsonnet score is removed the differences between the predictions of the *EuroSCORE* and the Bayes scores become more apparent.



The Bayes scores predict mortality much more accurately than the Parsonnet score and slightly better than the *EuroSCORE*.



Application of risk stratification to models of performance monitoring

Comparative performance tables

In the 1998 report we focussed on collating and analysing pre-operative data and exploring different risk stratification algorithms. Intentionally conspicuous by its absence was any form of unit related performance assessment. The reasons were threefold. Firstly, for this purpose the data would need robust validation; secondly, the aim of the exercise had been to demonstrate that downloading and merging of data for basic analyses was feasible. Finally, it became clear that the mathematics behind the development of fair and meaningful comparative methodologies was not trivial; it still remains an evolving science. Since then things have moved on. The advent of clinical governance as part of the 1997 NHS reforms has put benchmarking and comparative performance indicators high on the agenda, as evidenced by the public release of the Department of Health Clinical Indicators, the production of the Dr Foster Good Hospital Guide and the availability of the National Casemix Office's Performance Analysis Toolkit which allows any English Trust to compare itself with any other English trust using a variety of indicators. The DoH has now established an iterative process for reviewing its strategy on performance indicators and the next set of clinical indicators will be combined with Health Authority indicators. This will lead to a set of headline indicators at trust level, together with a more detailed benchmarking set of about 400 indicators for more in depth analysis. Many of these clinical indicators are based on data required for the implementation of the National Service Frameworks, which, in the case of coronary heart disease, includes the Society of Cardiothoracic Surgeons recommended dataset.

However, the public release of comparative performance indicators or league tables remains a sensitive issue among clinicians, largely for fear of misinterpretation by the media and patients. To counter these concerns some basic policy recommendations concerning the public reporting of outcome data have been summarised in the Introduction on page 14. In this section we examine in detail some of the more technical aspects of how comparative data can be presented in an easily interpretable fashion, so as to encourage rigorous, but fair, evaluation of clinical performance, whilst highlighting potential pitfalls in interpretation.

The most appropriate means of presentation is still a subject of protracted debate, and the current lack of well validated data provides an opportunity for us to illustrate some of the issues surrounding the presentation of potentially sensitive data. Although based on real data, the following analyses should be taken purely as illustrative, and are simply intended to stimulate discussion.

To this end, the Society are in collaboration with the Nuffield Trust, the RAND organisation and the California Department of Health to draw on the experience of the public release of outcome data in North America. Any additional feedback on ways in which the fairness and objectivity of the presentation might be improved would be welcomed by the Society.

Why do we need statistical analysis at all?

Statistical analysis attempts to allow for the effects of natural variability that arises from inevitable and unpredictable differences between patients and surgeons, as well as between operations and outcomes. Any analysis based only on a limited number of cases will always be subject to errors. If a surgeon operates on two patients and they both survive he has a mortality of 0%, but if one dies he has a statistical operative mortality of 50%. No one would believe that either percentage represented the real risk for patients operated on by that surgeon. Assuming that one of the first two patients died following their operation, the surgeon might then do another 48 operations with no deaths. After the fiftieth operation his mortality would be calculated as 2%. At this point, a reasonable person might guess that this surgeon's true mortality rate is around 2%, but would not have been prepared to make a guess at his level of performance based on the first two cases. So, the greater the number of cases with known outcomes the more certain one can be of the true risk of an operation under that surgeon. Statisticians cope with this by creating mathematical "confidence limits" around any outcome based on the number of observations or cases performed. In the example above where one of two patients dies we cannot be absolutely confident that that surgeon's mortality is 50% but we can be 95% confident that the actual risk lies between 2.7% and 97.3%. By the time the surgeon has done 50 cases with only one death, we can be 95% confident that his real operative mortality lies between 0.1% and 12.0%. So the greater the number of cases, the more accurate the assessment becomes.

In practical terms, we look at data over defined time periods: one year, two years or three years, so that there are a meaningful number of cases to analyse¹⁹. Even so, these short series represent only a snapshot in time, and may not necessarily be representative of overall performance; even if a surgeon or unit had no deaths within the chosen timeframe, we would not necessarily believe that this happy situation would continue forever. Similarly, a short run of bad outcomes should be regarded with caution¹⁹.



It follows that decisions regarding so-called 'divergent' outcomes must take into account random variability and should not be based simply on the observed rate. Casemix must also be considered.

Increasing the precision of the comparison

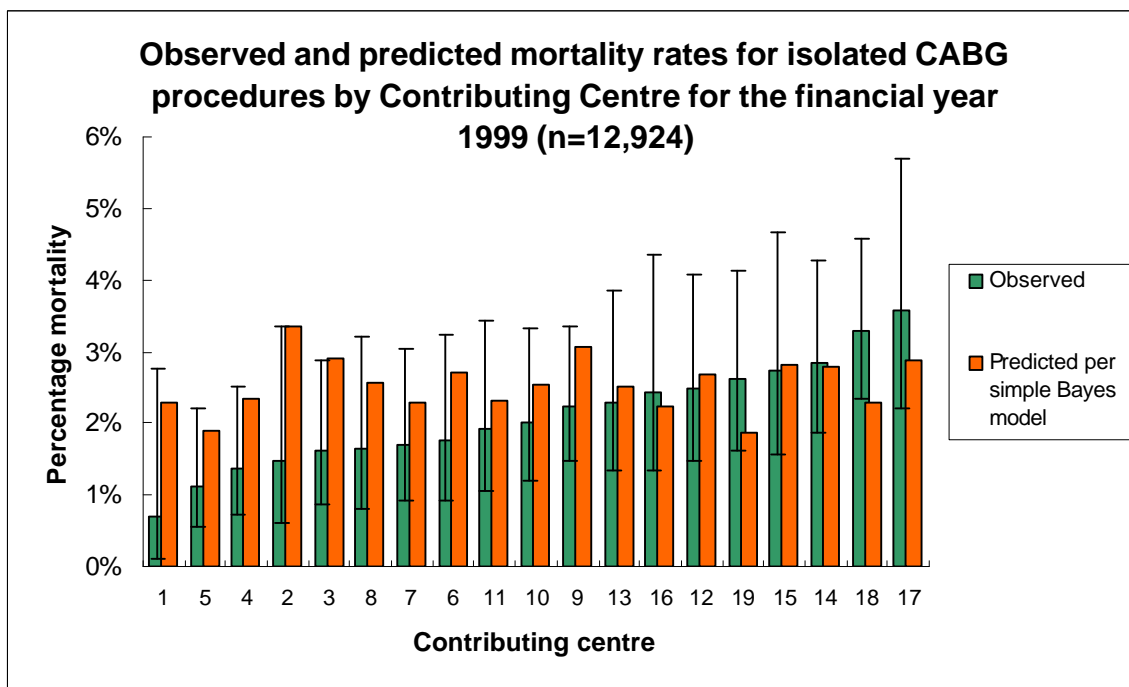
The precision with which comparisons can be made depends primarily on the ratio between the number of adverse outcomes and the total number of cases. Thus, the outcomes of a high risk procedure can be judged with greater statistical certainty using fewer operations than for a low risk procedure. Increasing the 'denominator', or total number of low risk procedures for analysis, may require:

- Aggregating the number of procedures over time: for example, three years' activity may be necessary to obtain enough observations for confident results.
- Combining well-defined, common groups of operations for analysis.
- Using 'near-misses' and deaths as adverse events instead of deaths alone.

Detecting divergence from benchmarks

Relevant presentation issues include:

- Graphical techniques for displaying data summaries are important. For example, when monitoring the performance of a single individual or institution over time, one might plot the cumulative excess of observed over predicted mortality^{26, 27} although care is required with assessing the statistical significance of any apparent divergence²⁰. These are described in detail in a later section (see pages 100 - 107).
- Risk adjusted outcomes can be related to observed or actual outcomes to provide 'indirect standardisation'. For example, if performance is measured by mortality, one can compute the standardised mortality ratio (SMR) or the difference from the expected mortality. However, it would be misleading to claim that statistical procedures can ever fully adjust for pre-existing risk factors.



The figure above shows the observed in-hospital mortality for isolated coronary surgery in nineteen different units using green bars, together with predicted mortality for each unit in orange. The vertical lines represent the 95% confidence limits around the observed mortality rate. For example, we can be 95% confident that the underlying mortality rate for surgery in unit 11 lies between 1% and 3.3%. The vertical lines for each unit tend to overlap those of other units, indicating that it is unlikely that there are any real differences between the different units' outcomes.



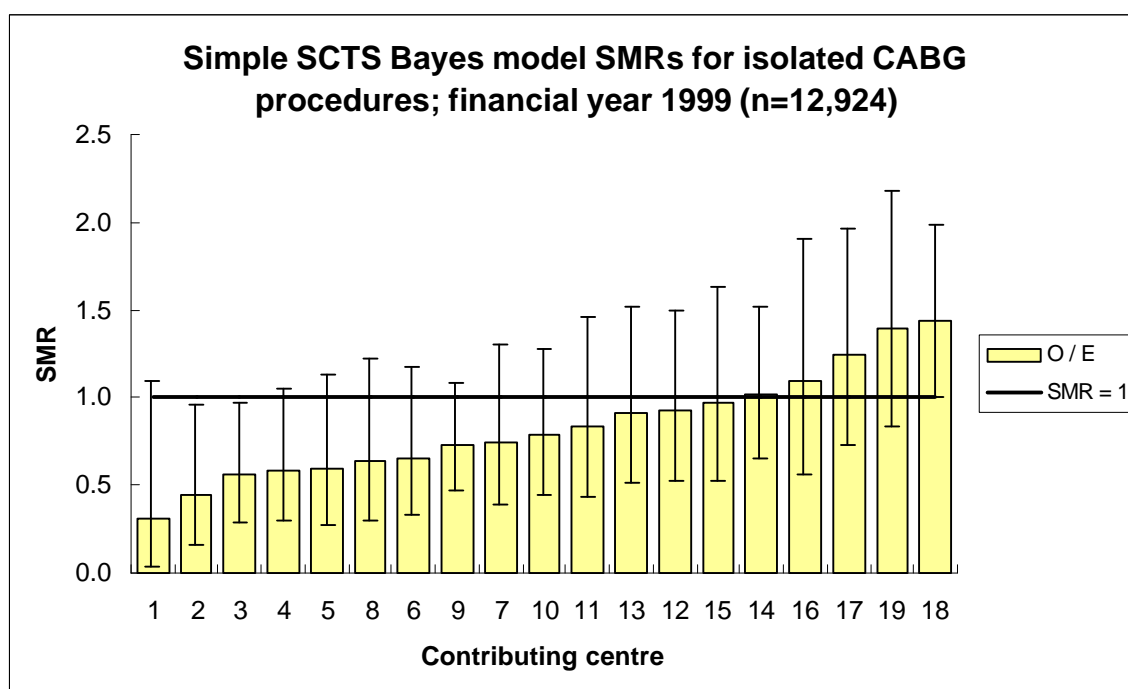
A popular method when comparing centres is to plot the observed performance and 95% confidence interval as used in NHS Clinical Indicators²¹. Sometimes a benchmark line is drawn - if the lower interval does not overlap that benchmark, then attention might focus on that centre. Thus, choosing a pragmatic and fair benchmark trigger is crucial if unfair conclusions are to be avoided.

Setting a benchmark 'threshold' for comparison

Benchmark thresholds may be either absolute or relative.

- Relative thresholds are defined according to statistical measures of overall performance, such as an average or two standard deviations from the average. Taken to its logical conclusion, this approach could penalise everyone above the benchmark at that time, even though there may be no real nor meaningful difference between their performance and the benchmark. Fifty percent of all people are below average intelligence, 50% of lawyers are below average as are 50% of doctors below average. This is a product of statistics and not a judgement on people, on lawyers or on surgeons.
- Absolute thresholds rely on specification of the upper end of 'acceptable' performance, and possibly the lower end of 'unacceptable' performance with a grey area that lies in between.

The figure below plots the standardised mortality ratio (SMR) for each unit. This is the ratio of observed to predicted mortality. This method compensates to some extent for casemix, but is highly dependent on the reliability of the method used to adjust for casemix. Note that the rank order of the units does not change dramatically.

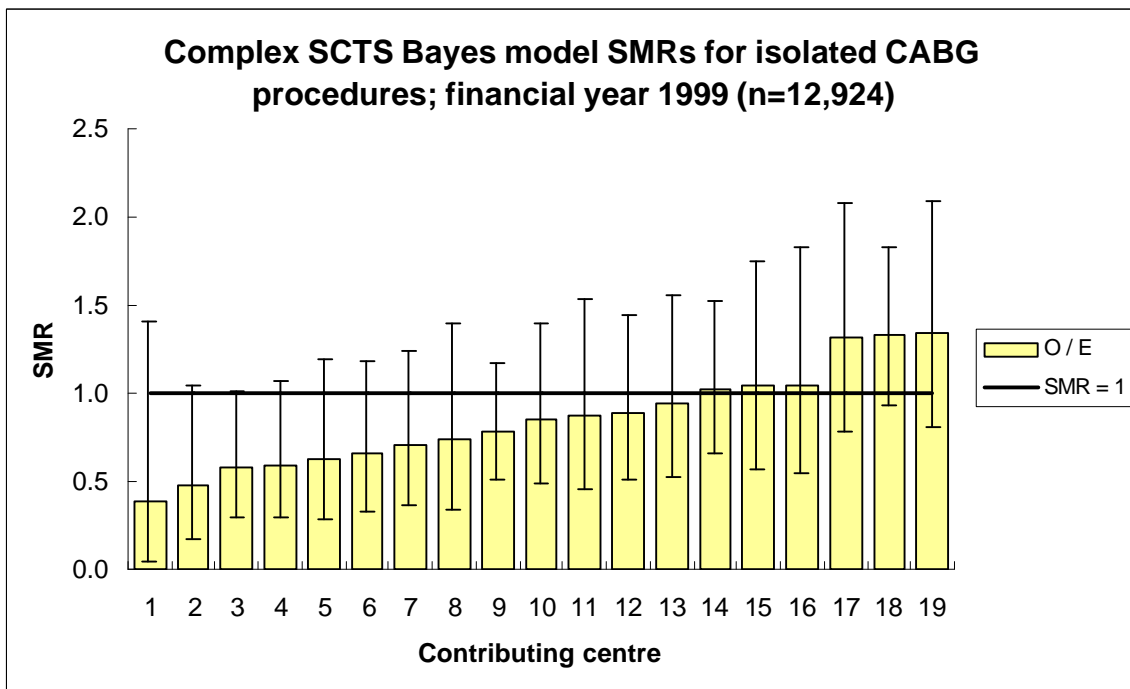
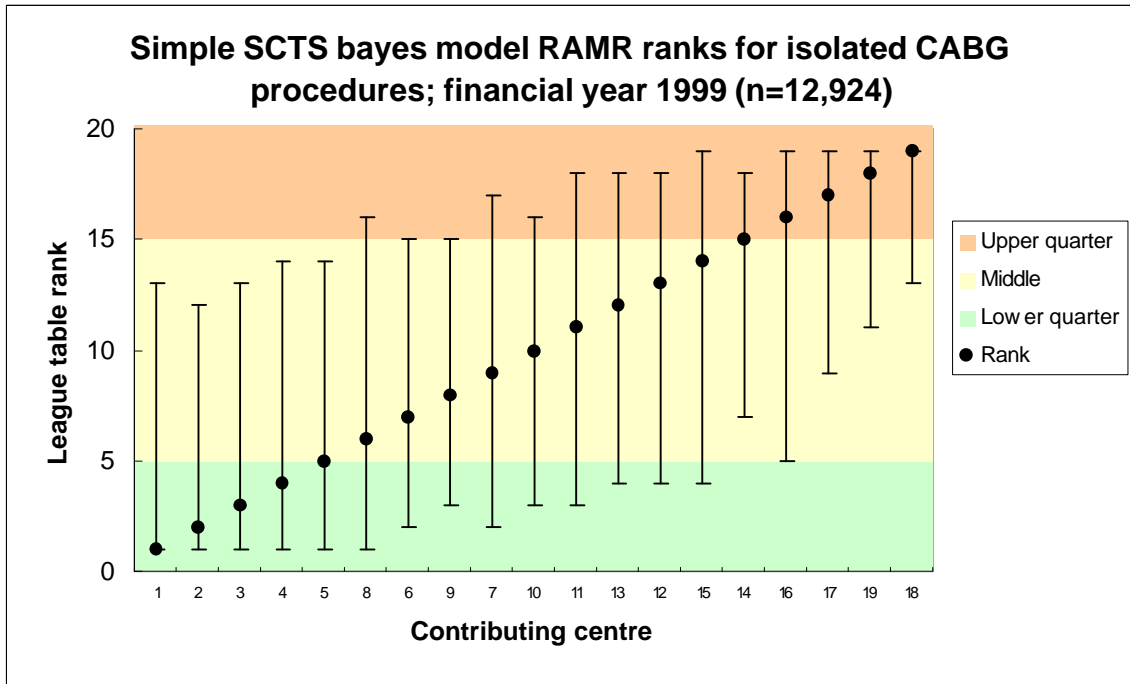


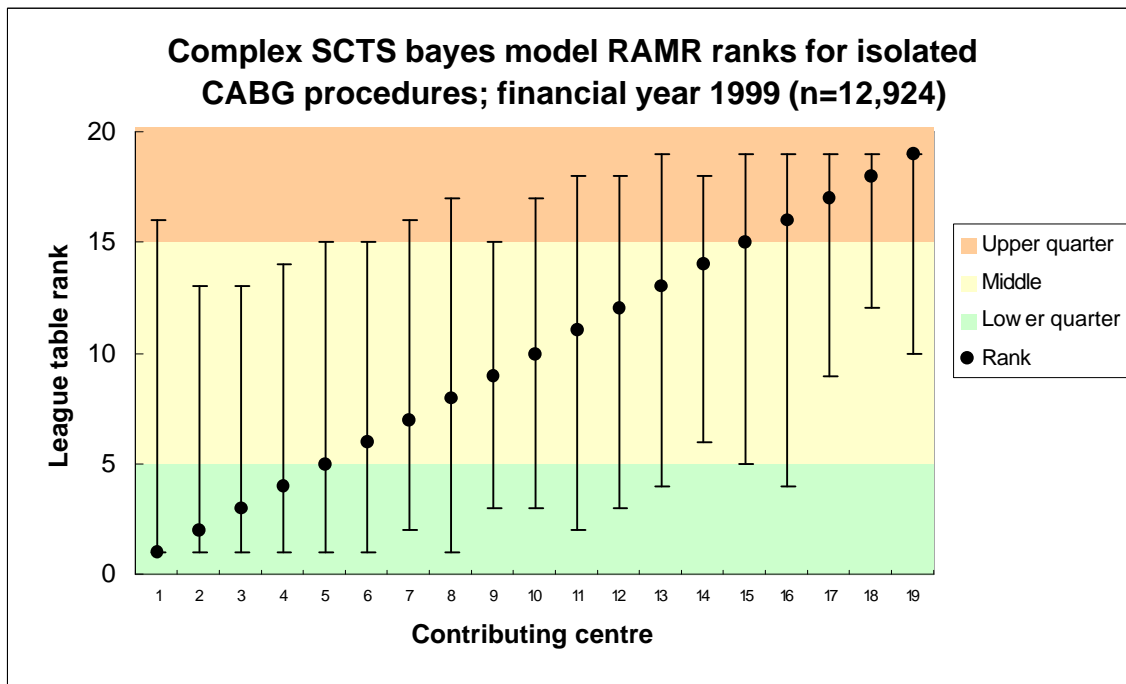
However, by chance alone, like any event that has a random component such as the toss of a coin, one in forty centres will be identified as 'significantly below average', even if their true performance were within the benchmark level. This indicates the need for caution in interpreting 'statistically significant' results. This technique is essentially testing the hypothesis that all the units have exactly the same underlying mortality rate which is highly unlikely in the real world.

In addition, the practice of ranking institutions to create 'league tables' has the potential to be extremely misleading, since ranks are notoriously sensitive to chance variability. For this reason, techniques have been developed for placing 95% confidence intervals around observed ranks²²; but this is not always enough. League tables of cardiac units in North America have clearly shown that a unit's position can vary substantially from one year to the next, more due to the effects of chance than to a change in the quality of care delivered by the unit^{23 19}.



The figure below illustrates this effect. It shows the ranks and 95% confidence interval determined according to the simple SCTS Bayes model for 19 centres, generating a typical pattern in which there is great uncertainty around the observed rank - we cannot be 95% confident that any particular unit lies in the top quarter. The same holds true when the more accurate complex SCTS Bayes model is used in the analysis on the same group of centres, but here the confidence limits are even wider, indicating even greater uncertainty around the ranks. This reflects the increased accuracy of the complex model over the simple model; it compensates for greater and more complex variations in casemix.





There is considerable danger in generating spurious 'false-positive' findings when carrying out many comparisons, although statistical techniques do exist for dealing with this phenomenon. One such technique is the Bonferroni adjustment, which essentially widens the confidence intervals that are used to indicate uncertainty, whereas the shrinkage estimation²⁴ method makes adjustments to all estimates of outcome rates based on an estimate of the expected regression-to-the-mean, pulling them towards the overall average. These methods could, of course, lead to an excess of 'false-negatives' in which genuinely divergent behaviour goes undetected.

The Statistician's phrase 'Regression-to-the-mean' describes the tendency for institutions that have been identified as 'extreme' to become less extreme when monitored in the future - put simply, part of the reason for their extremeness was a run of good or bad luck. This simple phenomenon could lead to spurious claims being made about the benefit of interventions to 'rescue' failing institutions. Shrinkage estimation is intended to counter this difficulty.

Perhaps the most important idea is that there is little gained from measuring variability in outcomes unless one can suggest underlying causes and remedial interventions. Someone must always be bottom of any 'league table', and the vital issue is whether they are truly divergent and, if so, why? Investigating the underlying reasons for variability in outcomes is not straightforward: while adjustment for case-mix is, in principle, possible, one must keep in mind that clinicians may respond to individual patient's situations in different but appropriate ways. Investigations must value that clinical skill.

Exploring options in the future

Subsequent reports will need to report such data in earnest, and it is vital that this is done in a way that balances the needs of the stakeholders. The data presented above are encouraging, but the data is not validated, so real conclusions cannot be drawn with any adequate degree of confidence. Validation techniques are being developed in conjunction with the Nuffield Trust in the UK, the RAND organisation in the US and the California Department of Health. This project, which includes 10 cardiac surgical units, is in the process of developing robust data validation methodologies, and will then go on to explore different ways of presenting outcome data that are meaningful and useful to both surgeons and patients alike.



Cumulative Sum (CUSUM) analyses

Standard CUSUM

The *standard CUSUM* chart plots a cumulative number of events against time²⁵. In the context of heart surgery, an event might be an adverse outcome, such as post-operative death, and time might be defined by the date-order in which operations were performed, the operative sequence. Most frequently, the adverse outcome in these kinds of analyses is an operative death. In order to standardise the time-frames, the first operation would be given the operative sequence number of 1, the second operation would be given a operative sequence number of 2 and so on.

The calculation of the *standard CUSUM* value is relatively simple. Each of the observed outcomes is given a numerical value: the adverse outcome is given a value of 1 and other outcomes are given a value of 0; these values are equal to the outcome rate for the individual patients. These values are then summed across the operative sequence to give the CUSUM value. In mathematical notation, this would be:

$$C_n = \sum_{p=1}^{p=n} \Delta_p$$

where

- n is the number of operations in the sequence
- C_n is the CUSUM value at operation number p ; the observed number of adverse outcomes
- Δ_p is the observed outcome for the patient at operation number p

These CUSUM values can be plotted against a predicted outcome rate, in order to visualise differences between observed and predicted outcome rates. To calculate the predicted CUSUM value the predicted outcome rate is first converted to probability. For a 5% risk, the probability of an adverse outcome would be 0.05, for a 10% risk the probability of an adverse outcome would be 0.10, and so on. These probabilities are summed across the operative sequence in much the same way as the observed outcome rates, but, for the purposes of charting, they are rounded down to the nearest whole number. For example, taking a series of 99 operations where each patient has a 1% risk of the adverse outcome, the corresponding probability of the adverse outcome would be 0.01 for each patient. Summing these probabilities over the series of 99 operations would give a predicted CUSUM value of 0.99. For charting purposes this would be rounded down to 0. If one more patient were added to the series, also with a 1% risk of the adverse outcome, the predicted CUSUM would then be $0.99 + 0.01 = 1.00$. On the chart, the predicted CUSUM value would now be 1. This explains the stepped appearance of both the observed and predicted CUSUM plots. In mathematical notation the predicted CUSUM would be:

$$P_n = \text{int} \sum_{p=1}^{p=n} d_p$$

where

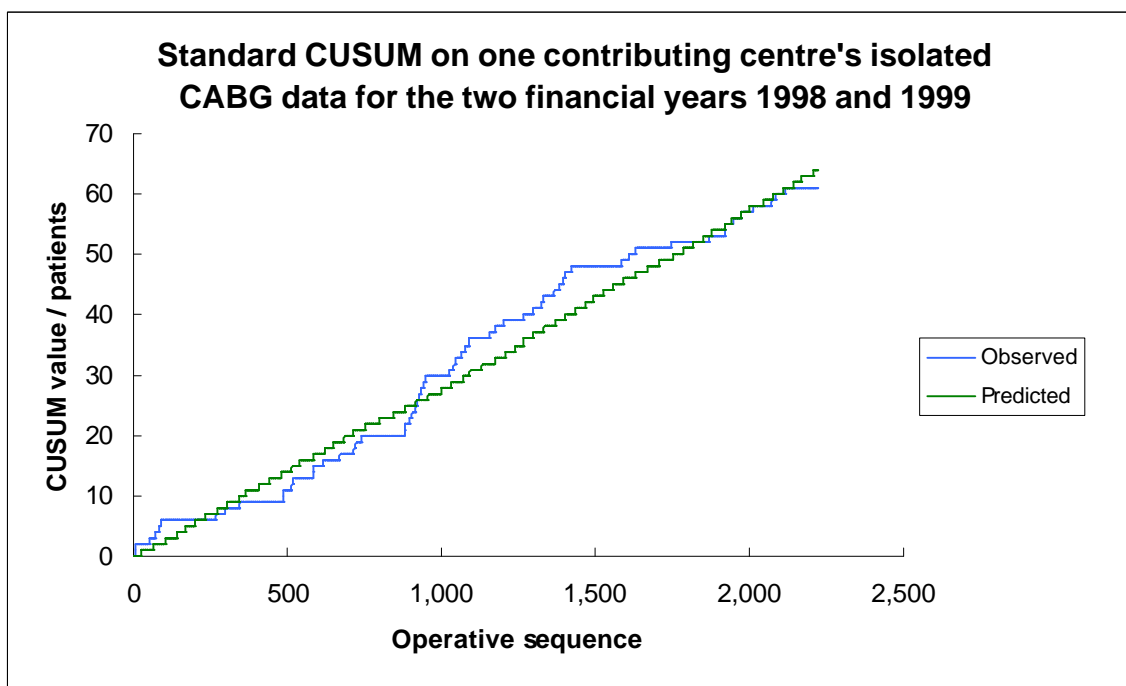
- P_n is the predicted CUSUM value at operation number p ; the predicted number of adverse outcomes
- d_p is the predicted outcome rate for the patient at operation number p

The following chart shows the observed and predicted cumulative number adverse outcomes for all isolated CABG procedures performed by one of the contributing centres over a two year period. An adverse outcome was defined as death for this particular chart. The cumulative risk across this series of 2,221 operations, according to the *Complex CABG Bayes score*, is 64 patients *i.e.*, 64 patient deaths are predicted.

It is obvious that the observed number of deaths is greater than the predicted number of deaths at some points along the curve, and less than the predicted number of deaths at other points. This is quite normal. The end-point of the curve shows that there were 3 patient deaths less than predicted. This sort of visual representation of data is sensitive to the point at which the analysis commences, and the point at which the



analysis ends. Had the curve stopped at operation 1,500 it would have shown that there were 5 more deaths than predicted. It is important to bear this in mind whenever examining these charts.



The results of this technique are heavily dependant on the risk model that is used as the predictor of the adverse outcome. Any risk score that has been designed to predict the adverse outcome of interest could be used in the calculations that create these curves. For example, in cardiac surgery risk scores such as the Parsonnet score, the *EuroSCORE* and Bayes scores are often employed.

In Parsonnet's original paper a Parsonnet score of between 0 and 4 carried an average operative mortality risk of 1%, a Parsonnet score of between 5 and 9 carried an average operative mortality risk of 5%, and so on up the risk score scale. Each time a surgeon operates on a patient with a Parsonnet score of between 0 and 4 there is, according to Parsonnet's original data, a 1% chance that the patient will die. If a surgeon operates on 100 such patients then Parsonnet would predict that one patient would die; the cumulative risk in this series is one patient death.

However, the predicted values quoted in Parsonnet's original paper are now out of date as practice in cardiac surgery has moved on, and, as a result, the observed mortality rate for each of the Parsonnet score groups has fallen. Any risk adjusted analysis of current mortality rates that employed the Parsonnet score as the predictor of death would tend to over-estimate the risk. In such an analysis, the observed mortality rate would tend to appear much lower than the predicted rate.

So, the exact position of the *standard CUSUM* curve with respect to the predicted number of adverse outcomes is greatly affected by the risk score that is used and the probability of death attached to that score. This applies equally to the *VLAD* and the *Risk Adjusted CUSUM* plots that follow. All of the following charts employ the *Complex CABG Bayes score* as the predictor of death, since this is a contemporary, risk score that has been shown to reflect recent outcome rates across the UK; in the case of the *Complex CABG Bayes score* the predicted risk is equal to the value of the score.

A note of caution: these visual representations provide useful tools to highlight trends that may need closer examination. In the format above, it cannot be used to determine a statistically significant departure from expected performance nor statistically significant differences between hospitals or surgeons. Furthermore, no risk score will be able to account for all the risk factors that may affect outcomes. Each will contain the most frequently encountered and measurable risk factors that impinge on the outcome they intend to predict. However, if only one surgically relevant risk factor that is absent from the risk model occurs frequently within the patient population of a hospital (or a surgeon) and not in the patient populations of other hospitals (or surgeons), then these plots will not paint an entirely accurate or fair picture. The *Complex CABG Bayes*



score is the best available approximation to the true risk at the present time. Even the *Risk Adjusted CUSUM* method should be used as an indicator and not as the final word in comparative mortality analyses.

Setting control limits around CUSUM curves

Using a *standard CUSUM* chart it is possible to identify deviations from a predicted outcome rate. It is useful to have some preset criteria for formally specifying that a deviation from a steady rate has taken place. These criteria are defined in statistical terms, and the choice of test is very important. Straightforward, standard tests of statistical significance after every operation are not appropriate. This would constitute “multiple testing”, and statisticians have determined that the results from these tests are misleading. “Sequential testing” techniques were developed during the 1940’s in an attempt to avoid these problems of multiple testing.

It is possible to draw two extra boundary lines on the *standard CUSUM* chart: an “alert” line and an “alarm” line. If the *CUSUM* line crosses either of these boundary lines there is an indication that the observed outcome rate has exceeded a pre-determined “target” rate, p_o . The mortality rate in the data used to generate the *CUSUM* chart above was 2.7%, so p_o might be set to 0.027. In order to avoid too many “false alarms”, a number of other parameters must be set: the first is a fixed chance that the *CUSUM* line will cross a boundary when the true outcome rate is still p_o ; this parameter is designated a , and is usually either 0.05, for a 5% chance, or 0.01, for a 1% chance. The next parameter is another outcome rate that is important to detect, which is designated p_1 ; it may be a higher outcome rate or a lower outcome rate, depending on whether the test is designed to look for worsening or improvement in performance. In the sample data used previously, the actual mortality rate was 2.7%, so a higher rate might be set at 3.7% and a lower rate might be set at 1.7%. Finally, the power of the analysis must be set, which defines the chance of correctly rejecting the “target” rate in favour of p_1 , according to the other criteria that have been set; this is denoted $1 - b$. The power of these analyses is typically set to 80%, which means that b would be, in probability terms, 0.20.

In order to calculate the positions of the boundary lines from these parameters, a number of components are needed:

$$a = \log \left[\frac{1 - b}{a} \right]$$

$$P = \log \left[\frac{p_1}{p_o} \right]$$

$$b = \log \left[\frac{b}{1 - a} \right]$$

$$Q = \log \left[\frac{1 - p_o}{1 - p_1} \right]$$

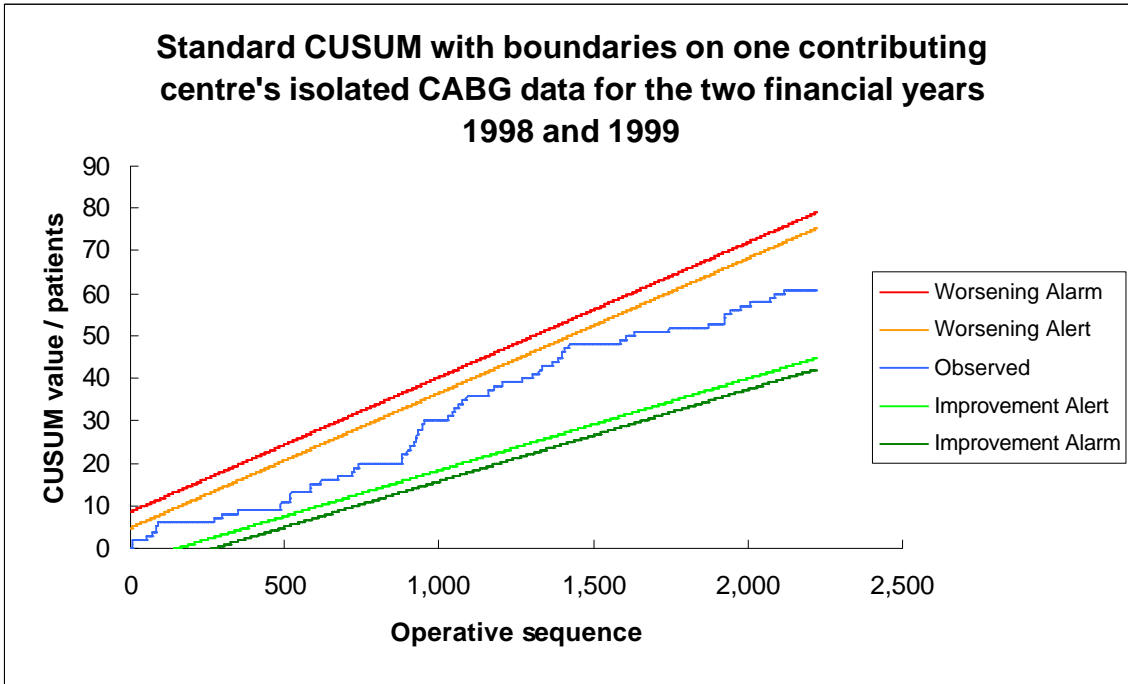
$$s = \frac{Q}{P + Q}$$

Finally, the boundaries can be calculated for each point along the curve; again n is the number of the operation in the sequence:

$$\text{Alert boundary:} \quad alert = n \cdot s - \left(\frac{b}{P + Q} \right)$$

$$\text{Alarm boundary:} \quad alarm = n \cdot s + \left(\frac{a}{P + Q} \right)$$

This chart shows that none of the boundary lines, as defined above, are crossed. This means that there is neither a worsening in the mortality rate nor an improvement. Different parameters would give different results.



This method simply takes a pre-defined outcome rate and determines whether or not the observed rate deviates from this rate. This method is not risk adjusted and therefore will not take account of any changes in casemix over time.



Variable Life-Adjusted Display (VLAD)

As with the *standard CUSUM* method, the *VLAD*^{26,27} could be used to examine any event rate as long as a suitable predictor exists for that event. For the purposes of the examples and charts in this section, an event was defined as death.

Plotting cumulative risk

This plot always has an upward trend and illustrates the cumulative risk over the operative sequence. Application and extension of this method can be used to calculate the cumulative risk associated with any group of patients using any suitable risk score. The following equation describes this approach:

$$r_n = \sum_{p=1}^{p=n} d_p$$

where

- p is the number of operations in the sequence
- d_p is the predicted outcome rate for the patient at operation number p
- r_n is the cumulative risk when the operative sequence is n

This kind of plot allows a visual comparison of predicted, cumulative risk from different sources, be they different hospitals or different surgeons. The curve is not shown, as it is not a necessary component of the *VLAD* analysis; the comparison of observed and predicted outcome rates is contained entirely in the *VLAD* itself.

The VLAD Plot

This curve attempts to provide a visual representation of performance against the predicted outcome rate. It may trend up or down over time.

Effectively, when the analysis begins the VLAD-value, which has units of patients, is set to zero. If a patient survives their operation, overall VLAD-value increases in a manner that relates to the patient's mortality risk. If the patient had a risk of 60% then the value increases by 0.6 patients, if the patient had a risk of 10% VLAD-value increases by 0.1 patients and so on. However, if the patient dies following the operation then the VLAD-value decreases. For the patient with a predicted mortality risk of 10% the decrease would be 0.9 patients, whereas for the patient with a risk of 60% the decrease would be only 0.4 patients. The death of a high risk patient is therefore appropriately weighted.

With each operation VLAD-value increases or decreases according to the risk of the operation, and at various times the net value will be positive, negative or zero. An overall positive value implies that more patients have survived than one would expect according to the risk model, whereas an overall negative value implies that more patients have died than expected. A zero value is exactly what the risk model predicts. Transitions from overall positive to overall negative values and back again are commonplace, and reflect the nature of normal surgical practice. The mathematical description of this concept is as follows:

$$P_n = \sum_{p=1}^{p=n} d_p - \Delta_p$$

where

- p is the number of operations in the sequence
- d_p is the predicted outcome rate for the patient at operation number p
- D_p is the observed outcome rate for the patient at operation number p
- R_n VLAD-value, the cumulative difference between the expected and observed outcome rates



An observed mortality rate can have a value of either 1 (certainty of death) or 0 (certainty of survival). If a patient has a 1% risk of dying then the value of d_p will be 0.01, if a patient has a 5% risk of dying then the value of d_p will be 0.05, and so on. If a surgeon operates on a patient with a mortality risk 1% and this patient survives then the positive value is:

$$0.01 \text{ (chance of death)} - 0 \text{ (survival)} = 0.01 \text{ patients}$$

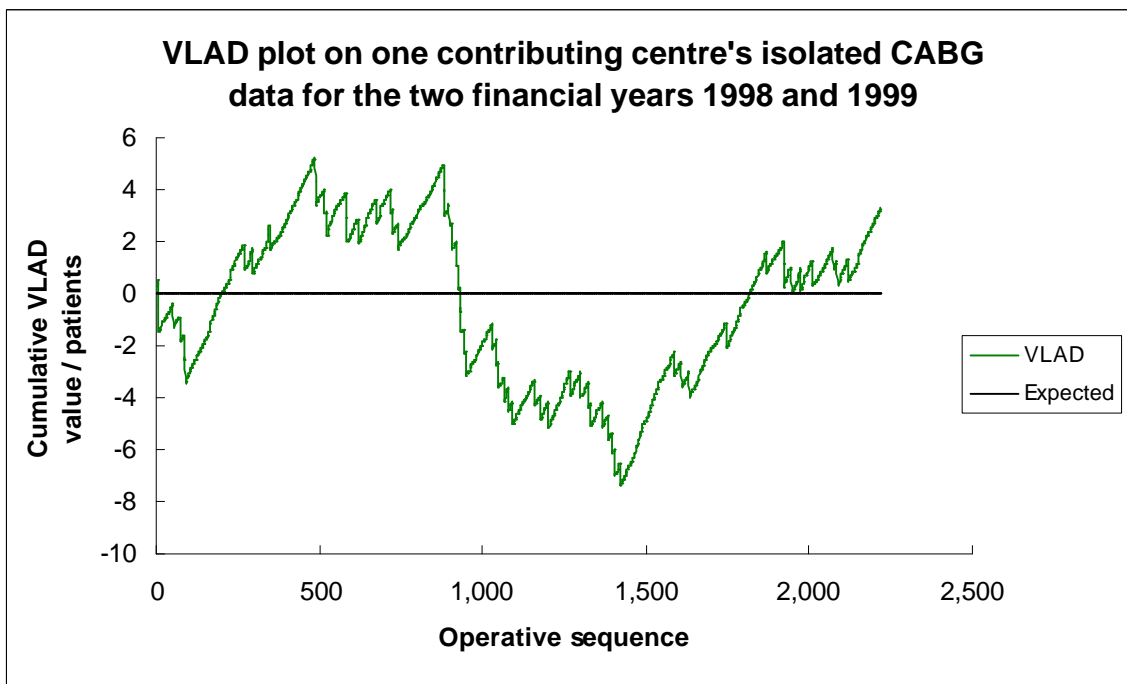
The VLAD-value for a series of 99 such patients would be 0.99 patients. Should the one-hundredth patient also survive then the VLAD-value is 1.00, or 1 patient fewer than expected. If, however, the one-hundredth patient dies, then the VLAD-value for this patient is:

$$0.01 \text{ (chance of death)} - 1 \text{ (death)} = -0.99 \text{ patients}$$

Adding this value to the cumulative VLAD value already determined for the previous 99 patients, all of whom survived, is 0.00. This is exactly the number deaths predicted by the risk model.

The **VLAD** shows, over the operative sequence, the number of patients deaths fewer than expected where the values are positive, or greater than expected where the values are negative. As with the CUSUM chart above, the chart below uses the **Complex CABG Bayes score** as the predictor of outcome for a group of isolated CABG patients from one of the contributing centres.

The end-point on this **VLAD** has an overall VLAD-value of 3.3 patients, or just over 3 patient deaths fewer than predicted. At various points along the curve, the overall VLAD-value rises as high as 5.2 patients and as low as -7.4 patients. The conclusions drawn from these analyses are, in part, dependant on the point in time at which the analysis starts and finishes. Had the analysis stopped at operation 1,500, then the conclusion would have been that 5 patients more than predicted had died, but another 700 operations later there are 3 fewer deaths than predicted. All these curves must be viewed with a critical eye.



In this format, it is not possible to determine statistical deviation from the predicted outcome rate. As with the **standard CUSUM** chart, it is important to remember that the shape of the curve and the end-points are dependant on the risk score that has been used in the calculations. An inappropriate or inaccurate risk score will produce misleading results.



Risk Adjusted CUSUM

The *standard CUSUM* method may be used to monitor changes in surgical outcome rates, but suffers in that it does not adjust for risk. This means that it may signal either an alert or an alarm simply because of an increase in the number of high risk patients having operations. The *VLAD* is risk adjusted, but this method has no robust methods for signalling deviations from the predicted outcome rate. A *risk adjusted CUSUM (RA-CUSUM)* method has been developed that comprises the best components of the *standard CUSUM* and the *VLAD*, and although the calculations associated with the method are a little more daunting than those used in the previous two methods, the underlying principles are the same. Those overwhelmed by the previous sections should skip this section.

The method utilises odds ratios for its calculations. As described previously in the section on Bayes modelling, odds ratios may be calculated as the probability of event over the probability of an alternate event. Therefore, using our previous notation, the odds on having an adverse outcome for an individual patient are calculated as:

$$\frac{d_p}{(1-d_p)}$$

Initially two hypotheses must be set: the first is H_O , which is the outcome rate that we wish to test. This could be set to the level of current practice. The second, alternative hypothesis, H_A defines the deviation from H_O to be detected. Either increases in the outcome rate or decreases in the outcome rate may be detected; a doubling in the outcome rate would set, a halving in the outcome rate or any other suitable differences may be chosen. There is an odds ratio associated with each of these hypotheses; the odds ratio associated with H_O is denoted OR_O , and the odds ratio associated with H_A is OR_A . Using our previous notation:

$$OR_O = \frac{d_O}{d_o} \text{ and } OR_A = \frac{d_A}{d_o}$$

where

d_O is the probability of the adverse outcome under H_O

d_A is the probability of the adverse outcome under H_A

If H_A is set such that a doubling in the rate is to be examined, d_A is clearly 2; if H_A is intended to look for a halving in the rate, d_A is 0.5. The method repeatedly tests H_O against H_A . Under H_O the odds on an adverse outcome for an individual patient, O_{pO} are:

$$O_{pO} = \frac{OR_O \cdot d_p}{(1-d_p)}$$

The corresponding probability of an adverse outcome for this same patient is:

$$d_{pO} = \frac{OR_O \cdot d_p}{[1-d_p + (OR_O \cdot d_p)]}$$

Under H_A the odds on an adverse outcome for an individual patient, O_{pA} are:

$$O_{pA} = \frac{OR_A \cdot d_p}{(1-d_p)}$$

And, the corresponding probability of an adverse outcome for this same patient is:

$$d_{pA} = \frac{OR_A \cdot d_p}{[1-d_p + (OR_A \cdot d_p)]}$$



A log-likelihood ratio can then be calculated for each patient; it is denoted W_n . The form of the calculation depends on the actual outcome for that patient. If the patient does not have the adverse outcome, the calculation is:

$$W_n = \log \left[\frac{1 - d_p + OR_O \cdot d_p}{1 - d_p + OR_A \cdot d_p} \right]$$

If the patient has the adverse outcome, the calculation is modified to:

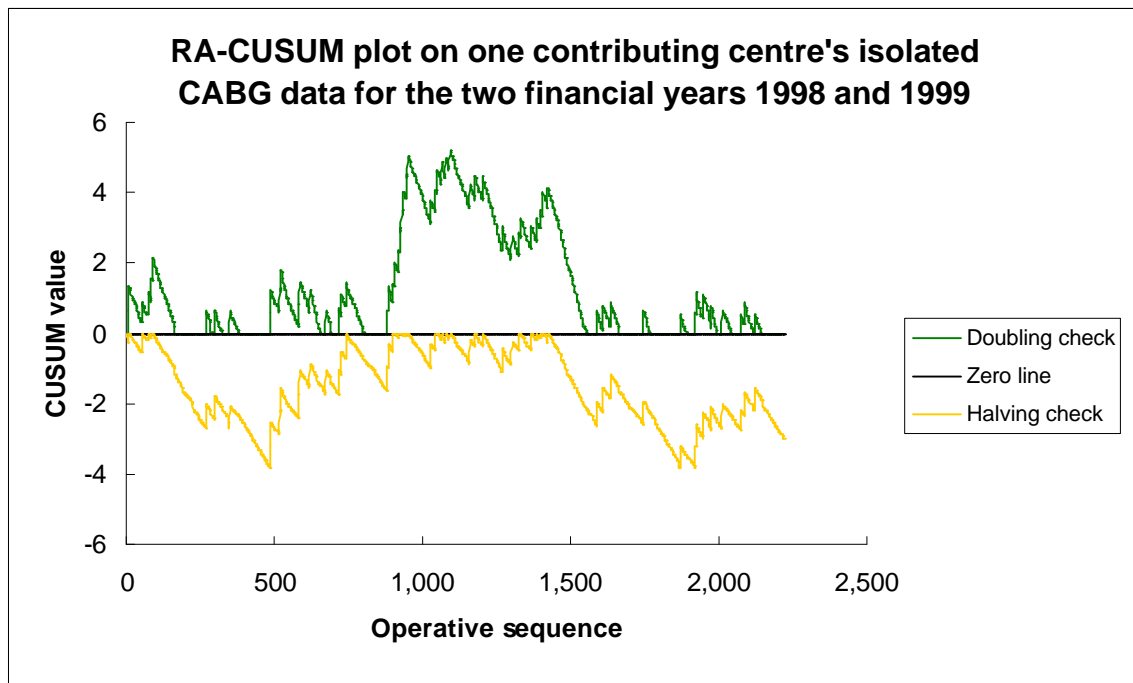
$$W_n = \log \left[\frac{(1 - d_p + OR_O \cdot d_p) \cdot OR_A}{(1 - d_p + OR_A \cdot d_p) \cdot OR_O} \right]$$

Essentially, the method plots a cumulative value against time, in exactly the same way as the previous two methods. The difference comes in the way that that value is calculated. A cumulative C -value is calculated for tests intended to look for deterioration in the outcome rate, and a Z -value for tests intended to look for an improvement. These values are described by the formulæ:

$$X_n = \max(0, X_{t-1} + W_n)$$

$$Z_n = \min(0, Z_{t-1} - W_n)$$

The “max” and “min” components of the formulæ simply mean that C_n is set to zero unless C_{n-1} was positive, in which case it is calculated from C_{n-1} and W_n , and Z_n is set to zero unless Z_{n-1} was positive, in which case it is calculated from Z_{n-1} and W_n .



In the chart shown above, an adverse outcome was set as a post-operative death and the probability of the adverse outcome was the *Complex CABG Bayes score*.

Alarm lines, parallel to the x-axis, can be drawn on the chart, to indicate when the observed value signals either a deterioration or an improvement in outcome rates. The calculation of the position of these alarm lines is complex, and beyond the scope of this report. It is as important to look for significant improvements in the outcome rate as it is to look for deteriorations, as good practice should be examined and disseminated.

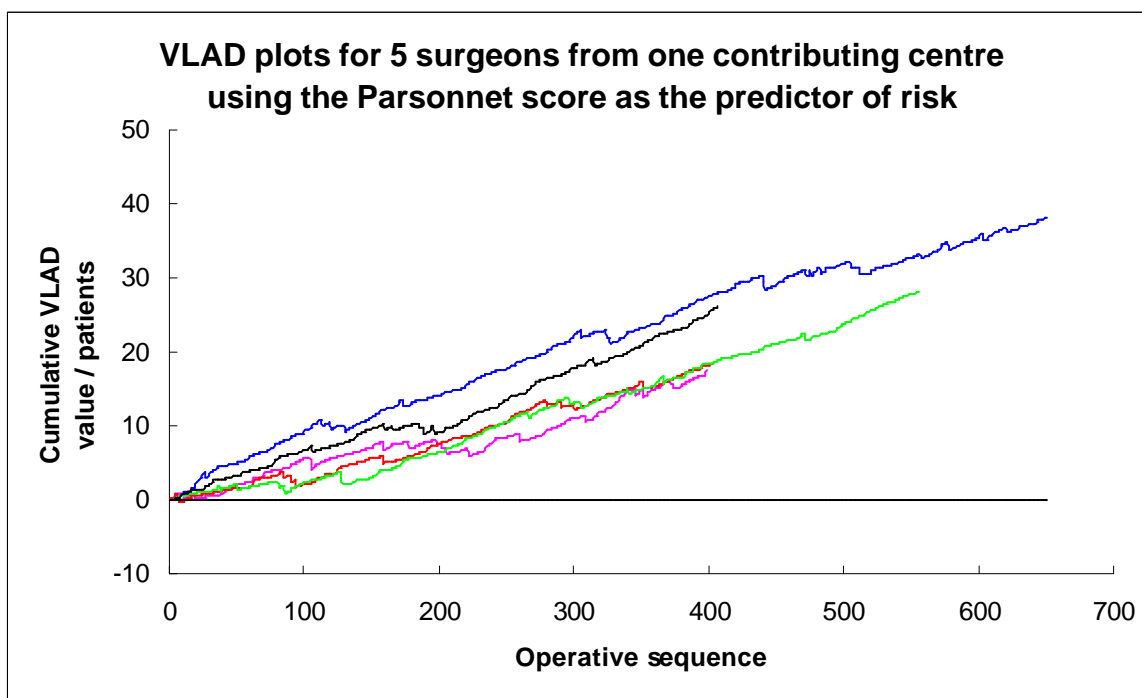


Beyond mortality by Mr Geoffrey Berg

Consultant Cardiothoracic Surgeon, Western Infirmary, Glasgow

Application of the VLAD technique to the outcome death and/or near miss of death

Mortality is a readily measurable end-point but a poor indicator of quality of care, cost-effectiveness or use of resources. Our present mortality data does not differentiate the average surgeon delivering stable patients from theatre from the poor surgeon who has the back up of a good intensive care unit clawing patients back from the brink. Using death following first time coronary artery surgery as a benchmark for a surgeon's performance, even with risk stratification, is not sensitive due to the overall low mortality rate.



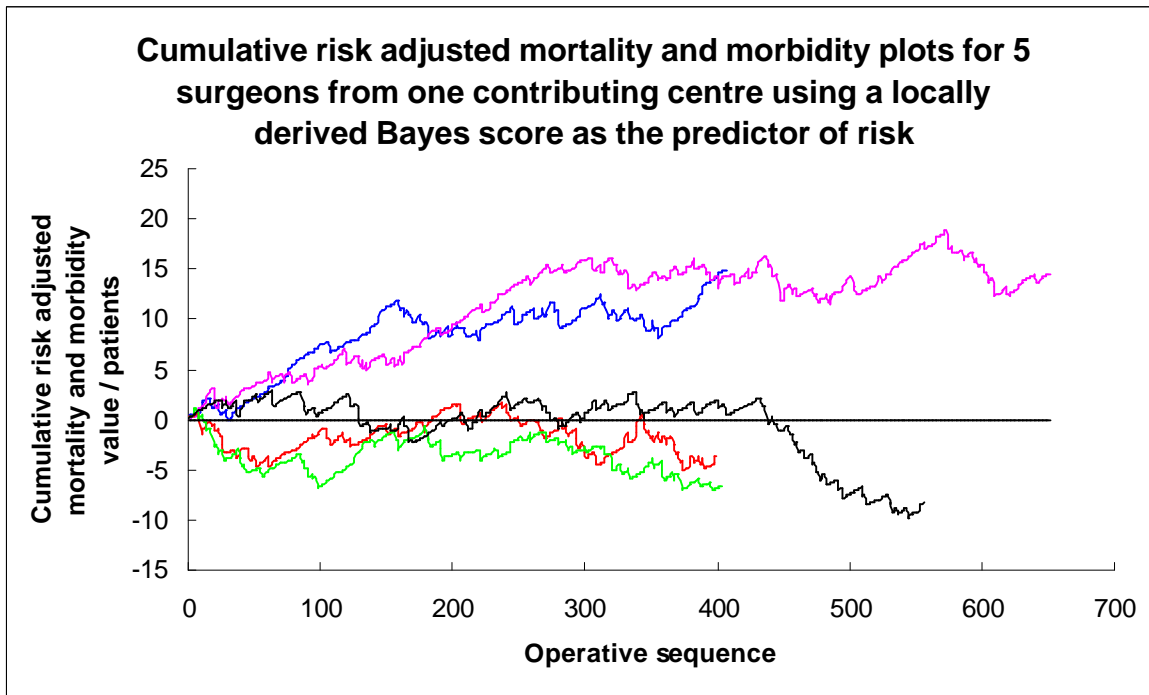
Morbidity following cardiac surgery is much more common, but it is more difficult to collect accurate data. Although many units are collating post-operative complication data, there are no requirements to publish these data. Numerous fixed scoring systems have been developed for cardiac surgery patients but there may be problems in applying these systems to different populations. Not only may patient demographics differ between countries but also within countries. A system developed in one time frame may not be accurate in a different time frame due to advances in surgical technique and post-operative management, or due to changes in resources. These systems also do not take account of local expertise.

For these reasons our unit developed a local system of risk stratification to be used in conjunction with established systems – to measure death and near miss of-death. To try and define an easy measurement of major morbidity following cardiac surgery we developed a series of measurements to record near miss of-death. This has been incorporated into a computer system so that the risk of death or near miss of-death can be estimated pre-operatively.

The development of various post-operative complications will often lead to prolonged ventilation or intensive care stay and initially we selected ventilation time, intensive care stay and time-to-discharge as a measure of near miss of-death. Although this incorporated many of the common post-operative complications such as stroke and peri-operative myocardial infarction, there were a number of important events that it missed. Intra-aortic balloon pump inserted post-operatively, acute renal failure, cardiac arrest and patient returned to theatre for a new cardiac procedure were added to ventilation greater than 72 hours, intensive care stay greater than 96 hours and discharge from hospital greater than 12 days.



If a patient had any of these events during the post-operative stay, they are classified as having a near miss of-death. This near miss event can then be counted and used as part of the audit process. Using Bayesian analysis, a scoring system can be developed and patients can be risk stratified for near miss of-death preoperatively. This system has demonstrated some interesting findings; less than 10% of our coronary patients have a near miss of-death but over 20% of our valve patients do. Nearly 30% of our coronary patients with poor pre-operative left ventricular function either die or have a near miss of-death. We have found that obese patients have a lower mortality than average but an average risk of near miss of-death. As shown below cumulative risk adjusted plots can be made and may be more sensitive in highlighting consultants' or trainees' performance.



It is possible to collect and audit post-operative morbidity using agreed definitions. All the units in Scotland submit their results to the Scottish Audit of Cardiac Surgery. Thirty one post-operative data fields are included (using CCAD definitions). The results are published and are made known to our purchasers and referring physicians. Patients are now aware that adverse events happen in hospitals but are also aware that many of them are preventable. If we want to improve our overall outcomes we must move on from measuring and reporting mortality and record and report morbidity, late mortality and even relief of symptoms.