How to evaluate and improve the quality and credibility of an outcomes database: validation and feedback study on the UK Cardiac Surgery Experience

Leon G Fine, Bruce E Keogh, Shan Cretin, Maria Orlando, Mairi M Gould, for the Nuffield-Rand Cardiac Surgery Demonstration Project Group

Abstract

Objectives To assess the quality and completeness of a database of clinical outcomes after cardiac surgery and to determine whether a process of validation, monitoring, and feedback could improve the quality of the database.

Design Stratified sampling of retrospective data followed by prospective re-sampling of database after intervention of monitoring, validation, and feedback.

Setting Ten tertiary care cardiac surgery centres in the United Kingdom.

Intervention Validation of data derived from a stratified sample of case notes (recording of deaths cross checked with mortuary records), monitoring of completeness and accuracy of data entry, feedback to local data managers and lead surgeons.

Main outcome measures Average percentage missing data, average κ coefficient, and reliability score by centre for 17 variables required for assignment of risk scores. Actual minus risk adjusted mortality in each centre.

Results The database was incomplete, with a mean (SE) of 24.96% (0.09%) of essential data elements missing, whereas only 1.18% (0.06%) were missing in the patient records ($P < 0.0001$). Intervention was associated with (*a*) significantly less missing data (9.33% (0.08%) $P < 0.0001$); (*b*) marginal improvement in reliability of data and mean (SE) overall centre reliability score (0.53 (0.15) *v* 0.44 (0.17)); and (*c*) improved accuracy of assigned Parsonnet risk scores (κ 0.84 *v* 0.70). Mortality scores (actual minus risk adjusted mortality) for all participating centres fell within two standard deviations of the mean score.

Conclusion A short period of independent validation, monitoring, and feedback improved the quality of an outcomes database and improved the process of risk adjustment, but with substantial room for further improvement. Wider application of this approach should increase the credibility of similar databases before their public release.

Introduction

Public release of health outcomes is often proposed as one mechanism for holding healthcare providers accountable for the quality of care they deliver. The single most important concern about releasing healthcare information to the public is that the data for assessing outcomes adjusted for case mix may be biased or incomplete, leading to flawed conclusions and thereby undermining the credibility of national programmes.

Cardiac surgeons in the United Kingdom have been at the forefront of data collection on clinical outcomes, and the Society for Cardiothoracic Surgeons of Great Britain and Ireland (SCTS) has published annual reports on mortality from cardiac surgery that have protected the anonymity of the surgical centres.[1] This year the SCTS went one step further and published unadjusted mortality for isolated coronary artery bypass surgery and aortic valve surgery for all units in the United Kingdom on its website[2] and in its 2000-1 annual report.[3] While expressing reservations over the value of reporting unadjusted or inadequately adjusted outcomes,[4 5] the SCTS felt unable to proceed to full risk adjustment because of concerns about the quality and completeness of data on each patient within its national database.

The present project was undertaken to address these concerns and to determine what needs to be done to improve a database to the point that participating centres and government bodies will be comfortable with full disclosure of outcomes by centre of origin. To this end we launched a demonstration project, the goal of which was to assess the reliability and completeness of the existing national outcomes database. Concurrent with a baseline assessment, we instituted a short programme of validation, monitoring, and feedback in an attempt to improve data quality.

Methods

Conducted in partnership with the SCTS, the project was part of collaboration on public release of information about the quality of health care between the Nuffield Trust in Britain and the RAND Health Program in the United States.[6] In this project we focused on only one outcome measure—risk adjusted mortality in hospital after isolated, first time coronary artery bypass surgery. We chose mortality in hospital in preference to 30 day mortality because this had been fully debated and agreed by the SCTS. The decision was reached on the basis that the former is more clinically relevant, is easier to validate, and is used

Department of Medicine, Royal Free and University College Medical School, University College London, London WC1E 6JJ
Leon G Fine *professor and head*
Mairi M Gould *project coordinator*

Department of Cardiothoracic Surgery, Queen Elizabeth Hospital, Birmingham B15 2TH
Bruce E Keogh *consultant surgeon*

RAND Health, PO Box 2138 Santa Monica, CA 90407-2138, USA
Shan Cretin *senior scientist*
Maria Orlando *associate behavioural scientist*

Correspondence to:
L G Fine
l.fine@ucl.ac.uk

BMJ 2003;326:25–8

Members of the project group are listed on bmj.com

internationally (such as by the US Society of Thoracic Surgeons, New York State Society of Surgeons, and Pennsylvania Society of Thoracic Surgeons). The prerequisites for participation were that participating centres should agree to release the findings of the study for publication, including the outcome in each participating centre by name, and that they were able to capture data in an electronic format.

Design of the project

The specific questions we addressed were:

- Was there a substantial amount of missing data in the SCTS database that would make calculation of risk adjusted mortality unreliable?
- Could the data elements entered into the database be validated by review of patients' records?
- Could recorded deaths be substantiated by referring to mortuary records?
- Could the accuracy of and completeness of the database be improved after a period of monitoring, validation, and feedback in participating centres?

Developing a "gold standard"

We identified 17 essential data elements required to assign a preoperative severity (risk) score using the Parsonnet score[7] and the EuroSCORE (European System for Cardiac Operative Risk Evaluation).[8] A single study coordinator (MMG) was trained to become the "gold standard" for assessing the accuracy of scoring of these elements. Standard, internationally agreed definitions of each risk variable, as agreed by the SCTS, were used.[9] The study coordinator reviewed a sample of three case records from each of 10 centres, which had been coded by the operating surgeon, and resolved any discrepant codes for individual data elements with the help of an experienced cardiac surgeon (BEK). Re-scoring of the same records in a blinded fashion by the study coordinator about two months later revealed almost precise agreement (less than one discrepancy per patient record).

Retrospective validation, monitoring, and feedback

The 10 participating centres submitted data for the period from 1 April 1997 to 31 March 1998, providing 7711 cases of isolated coronary artery bypass surgery (anonymised with respect to patients). There was substantial variability between centres in how data were collected, from handwritten entries on to locally created forms (with subsequent transfer to a computerised database) to direct entry into an electronic database. There were also differences in the software packages used, although centres were required to export data in Microsoft Excel format.

We stratified the cases for severity using EuroSCOREs and randomly sampled about nine case records in each of six predefined risk strata from each centre—that is, about 54 case records per centre (since some centres had only a small number of cases in the highest risk stratum, the number sampled was less than nine in these instances). This was a key element of the validation strategy, which ensured that a full spectrum of case severity was examined in the validation process. The number of case records sampled provided 80% power to detect one standard deviation difference between the submitted and re-coded risk scores. The study coordinator re-abstracted the stratified random sample of records from each centre and re-coded the 17 essential data elements from each record. The data elements submitted by the centres were then compared with the re-coded elements, as were the overall risk scores calculated from the submitted and re-coded data.

We assessed the 17 essential data elements for completeness and reliability. We calculated the "percentage missing data" for each centre as the average percentage of data elements missing across all patients in the centre. We expressed the "average κ" for each centre as the average of the κ coefficients calculated for each of the 17 data elements. The κ coefficient measures reliability of scoring by two observers, adjusting for agreement by chance alone (κ=0 represents no agreement, κ=1 represents perfect agreement, and scores of 0.8 or better are considered "almost perfect"). For the purposes of this study, we computed an additional "centre reliability score" to reflect both reliability and completeness by multiplying the average κ for a centre by the proportion of data elements with sufficient information to calculate κ (so if κ was calculated for all 17 data elements in a centre, the centre's reliability score would equal its average κ).

Validation based on the previously developed definitions required that specific information be present in the patient record to allow a score to be assigned to a particular data element.

The project coordinator made at least two visits to each centre to provide feedback and monitoring on site. We also communicated with the centres by email and telephone. When misunderstanding about scoring particular data elements was widespread, we sent a memorandum of clarification to all centres.

Since mortality in hospital was the only outcome measure, we compared the mortality data submitted by each centre, using a "snapshot" view covering a one month period, with the records in the respective hospital mortuaries for that month and the subsequent three months. We adopted this approach because all mortuaries lacked an electronic database, making an electronic comparison of the data impossible.

Prospective phase

Having completed the retrospective analysis, we conducted a similar prospective exercise on a total of 2683 submitted patient records for the period 1 July to 30 November 2000, and re-abstracted a stratified sample of 430 records. Before this phase, two data items were modified in keeping with the SCTS definitions: dyspnoea (measured according to the New York Heart Association score) and angina (measured by the Canadian Cardiac Society score) were assessed as the most severe within two weeks before surgery. All other data codings remained unchanged.

The mortuary checks were carried out on only five centres in this phase.

Results

Completeness and reliability of the database.

A retrospective review of the original 12 month published database showed it to be incomplete, largely because of failure to transfer information from patient records into the database. In the sample of re-abstracted records combined across the nine centres that participated in both phases of the study, a mean (SE) of 1.18% (0.06%) of the essential data elements were missing, compared with 24.96% (0.09%) missing in the submitted data for these same records. This dif-

Percentage missing data and reliability scores for nine participating surgical centres which submitted mortality data for coronary artery bypass surgery for the 12 month retrospective phase and the five month prospective phase. Values are means (standard errors)

	Retrospective phase			Prospective phase		
	% missing data elements	κ Coefficient	Centre reliability score	% missing data elements	κ Coefficient	Centre reliability score
Submitted data	24.96 (0.09)	0.67 (0.11)	0.44 (0.17)	9.33 (0.08)	0.78 (0.06)	0.53 (0.15)
Re-abstracted data	1.81 (0.06)			4.04 (0.05)		

ference in percentage missing data was highly significant ($P<0.0001$). Mean (SE) reliability (κ) of the sample across the nine centres (confined to those data elements present in submitted and re-abstracted data) was 0.67 (0.11) (see table). However, when we adjusted this for the proportion of κ coefficients able to be calculated in each centre the average overall centre reliability score was only 0.44 (0.17).

One centre did not submit data for the second phase because of failing to capture the data in electronic format, leaving nine centres in the prospective phase. Compared with the retrospective analysis, the percentage of submitted data that was missing in the prospective phase had fallen significantly to 9.33% (0.08%) ($P<0.0001$), but the percentage of data missing in the re-abstracted records (4.04% (0.05%)) was still significantly lower than that of the submitted data ($P<0.0001$). Reliability of data elements present in both the submitted and the re-abstracted database also improved, although not significantly, resulting in a mean κ coefficient of 0.78 (0.06) (see table). The mean overall centre reliability score, adjusted for the proportion of κ coefficients able to be calculated in each centre, was 0.53 (0.15). Thus a tangible overall improvement had been achieved over the five month period of monitoring, but the final centre reliability score remained only moderate.

Calculation of risk scores

We calculated Parsonnet scores and EuroSCOREs, reflecting the severity of the case mix, for the submitted and the re-abstracted data. In the retrospective phase the overall submitted Parsonnet scores (but not EuroSCOREs) were marginally but significantly higher than the scores recalculated from the submitted data and from re-abstracted data (fig 1). In the prospective phase the submitted Parsonnet scores were more reliable than in the retrospective phase (κ 0.84 *v* 0.70), but coding of some of the data elements required for calculation of the EuroSCORE remained problematic and κ remained essentially the same (0.65 *v* 0.61).

Validation of deaths using mortuary records

Mortuary records were uniformly poor in all centres. Handwritten entries, which often failed to include the hospital number, were the rule. We identified isolated discrepancies between the one month sample of mortality records submitted by each centre and the data entered in the mortuary records. This was true for both phases of the study. Investigations at the centres revealed problems with computer software, particularly where data were transferred across different software systems. Given the low death rates (about 3%) isolated discrepancies such as these can substantially affect outcomes.

Observed and risk adjusted mortality

Figure 2 shows the mortality outcomes by centre for submitted data in the retrospective review. We found a high correlation between risk adjustment by Parsonnet score and that by EuroSCORE. The mortality scores for all participating centres fell within two standard deviations of the mean score. Because we could not confirm mortality data in the prospective study for all centres, we did not perform this analysis for the prospective data (see below).

Post-project check

At the formal closure of the project, we made a confirmatory check of the numbers of cases and deaths. Only five out of nine centres responded within four months by resubmitting their data. In two centres we confirmed the originally submitted numbers, but in the other three we found small differences in the number of cases, and the originally submitted numbers of deaths were lower than the confirmed numbers by 2/412, 2/333, and 5/287. These errors alter mortality calculations substantially.

Discussion

A simple evaluation of the completeness and reliability of a national outcomes database revealed that it was both incomplete and unreliable in some respects. After five months of validation, monitoring, and feedback it improved measurably but still left substantial room for further improvement. We identified ongoing errors in data transfer. Since we did not perform the exercise on centres where there was no intervention, we cannot quantify the degree to which the intervention contributed to the observed improvement.

A strength of the project was the complete independence of the study coordinator, who had no vested interests in any of the outcomes. The voluntary involvement of the SCTS in the exercise and the willingness of participating centres to have their outcomes

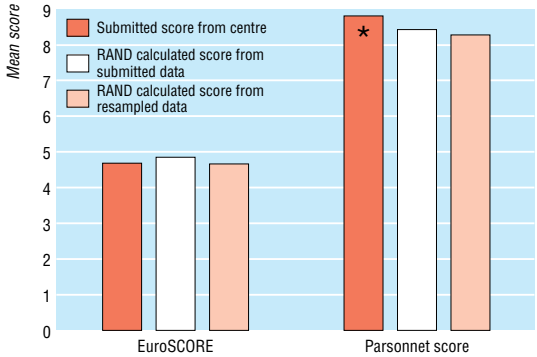

Fig 1 Average EuroSCOREs and Parsonnet scores for retrospective review of mortality data for coronary artery bypass surgery, comparing scores submitted by surgical centres, the same submitted data recalculated by RAND statisticians, and the scores calculated on re-abstracted data. Parsonnet scores submitted by centres were significantly higher than recalculated scores or scores derived from re-abstracted data (*P<0.05)

What is already know in this topic

Release of healthcare outcomes into the public domain has altered referral patterns and has led to improvement in some centres and elimination of others

The tacit assumption is that such outcomes data are accurate and can be relied on by the public and by healthcare providers to guide improvements

What this study adds

Sampling of a published national cardiac surgery database in England revealed it to be both incomplete and unreliable in its ability to yield accurate, risk adjusted outcomes data

An independent short process of monitoring, validation, and feedback improved the quality of the database

Such databases probably require an ongoing process of monitoring in order to allow data of adequate quality to be generated for the purpose of improving healthcare outcomes

published was seen to be a positive step toward continuing improvement in the quality of the society's outcomes database, making public release more credible.

Our study shows that the task of perfecting a database is never complete. It therefore makes sense to superimpose a permanent cycle of external monitoring and validation on all major outcomes databases. There is also an obvious need for improvement and tighter controls in information technology systems. The fact that a post-project check of the submitted data in the second phase of our study exposed inconsistencies in submitted information indicates that, at a minimum, all information released for publication should be subjected to an independent check before release.

In order to extend the experience gained in this demonstration to all cardiac surgery centres in the United Kingdom, we believe that the agreed SCTS minimum dataset should be adopted in each country and ultimately incorporated into institutional information technology systems. Over the course of the next year the clinical database in each cardiac unit in England will be linked to a central cardiac audit database administered by the NHS Information Authority and linked to the Office of National Statistics. This will allow long term tracking of mortality of all patients who have undergone cardiac surgery in those units and will enable us to understand who will benefit most from which operation.

To be effective, this endeavour must be coupled with an independent, comprehensive data validation process. Only when this is in place will patients have access to genuinely meaningful information and will healthcare providers be confident that they are being represented and judged fairly. This is particularly important given the recommendations of the Bristol Royal Infirmary Inquiry[10] and the subsequent agreement between government ministers and the Society of Cardiothoracic Surgeons to proceed towards the public release of individual surgeons' outcomes in England. It is also right and proper that the same processes of data validation and outcome presentation are uniform, or are uniformly adopted across the increasingly devolved healthcare systems of the United Kingdom.

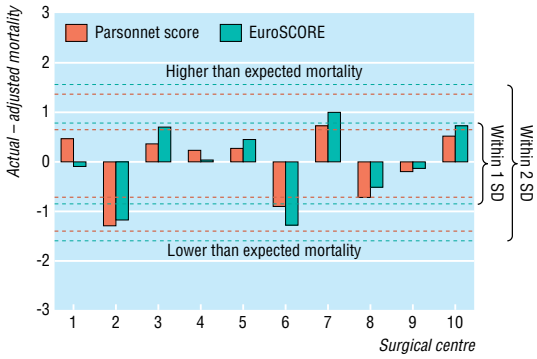

Fig 2 Actual mortality minus risk adjusted (expected) mortality by surgical centre in retrospective review. Risk adjustment was made with either Parsonnet scores or EuroSCOREs. Participating centres were 1 Queen Elizabeth Medical Centre, Birmingham; 2 Bristol Royal Infirmary; 3 Papworth Hospital NHS Trust; 4 Victoria Hospital, Blackpool; 5 Wythenshawe Hospital, Manchester; 6 Freeman Hospital, Newcastle; 7 Guy's and St Thomas's Hospitals, London; 8 Kings College Hospital, London; 9 Imperial College School of Medicine, Hammersmith Hospital, London; 10 University College London Hospitals, London

We are indebted to the following data managers, without whom this work could not have taken place: Miles Curtis, Vivienne Barnet, Karen Jack, Phillip Townbe, Paul Dillon, David Finch, Paula Clark, Valerie McLannahan, Sheila Jamieson, Joe Omigie and for their participation. We also thank members of the Cardiac Surgery Steering Committee, Jules Dussek, Kathy Rowan, Nick Black, Peter Walton, Robin Kinsman, and Tom Treasure, whose wealth and breadth of experience proved invaluable in guiding the project. Finally, we thank John Wyn Owen, secretary of the Nuffield Trust, and Robert Brook, director of the Rand Health Programme, for their creative ideas, constructive criticisms, and constant support.

Contributors: LGF was principal investigator and is guarantor of the work. BEK was co-principal investigator of the study and represented the Society of Cardiothoracic Surgeons of Great Britain and Ireland. MMG was project coordinator. SC and MO were the study statisticians. Other members of the Nuffield-Rand Cardiac Surgery Demonstration Project Group (listed on bmj.com) contributed to the design and execution of the study in collaboration with the above individuals.

Funding: This project was supported by a grant from the Nuffield Trust. The Nuffield Trust had a representative on the steering committee but played no role in determining the final design of the study or in the analysis and presentation of the results.

Competing interests: None declared.

1 Keogh B, Kinsman R. *National adult cardiac surgical database report*. London: Society of Cardiothoracic Surgeons of Great Britain and Ireland, 1999.
2 SCTS (Society of Cardiothoracic Surgeons of Great Britain and Ireland). www.scts.org (accessed 9 Oct 2002).
3 Keogh B, Kinsman R. *National adult cardiac surgical database report 2000-2001*. London: Society of Cardiothoracic Surgeons of Great Britain and Ireland, 2002.
4 Keogh B. Facts of life the figures can hide [viewpoint]. *Times* 2002 Nov 19.
5 Keogh B, Dussek J, Watson D, Magee P, Wheatley D. Public confidence and cardiac surgical outcome [editorial]. *BMJ* 1998;316:1759-60.
6 Marshall M, Shekelle P, Brook R, Leatherman S. *Dying to know: public release of information about quality of health care*. London: Nuffield Trust, 2000.
7 Parsonnet V, Dean D, Bernstein A. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation* 1989;79(suppl):3-12.
8 Roques F, Nashef SAM, Michel P, Gauducheau E, De Vincentiis C, Baudet E, et al. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. *Eur J Cardiothorac Surg* 1999;15:816-23.
9 Society of Cardiothoracic Surgeons of Great Britain and Ireland. National cardiac surgical database. Minimum surgical dataset definitions. www.scts.org/mindata97.html (accessed 9 Oct 2002).
10 Bristol Royal Infirmary Inquiry. The inquiry into the management of care of children receiving complex heart surgery at the Bristol Royal Infirmary. www.bristol-inquiry.org.uk (accessed 9 Oct 2002).

(Accepted 4 October 2002)