THE NUFFIELD TRUST–RAND COLLABORATION

IMPROVING THE CREDIBILITY OF INFORMATION ON HEALTH CARE OUTCOMES

The Cardiac Surgery Demonstration Project

RAND

The Nuffield Trust
FOR RESEARCH AND POLICY
STUDIES IN HEALTH SERVICES

THE NUFFIELD TRUST–RAND COLLABORATION

IMPROVING THE CREDIBILITY OF INFORMATION ON HEALTH CARE OUTCOMES

The Cardiac Surgery Demonstration Project

Leon G Fine, Bruce E Keogh, Maria Orlando, Shan Cretin, Mairi M Gould
Edited by Caroline White

RAND

The Nuffield Trust
FOR RESEARCH AND POLICY
STUDIES IN HEALTH SERVICES

ISBN 1-902089-90-1

Published by The Nuffield Trust
59 New Cavendish Street
London WIG 7LP

Telephone: 020 7631 8450
Facsimile: 020 7631 8451

E-mail: mail@nuffieldtrust.org.uk
Website: www.nuffieldtrust.org.uk

Charity Number: 209201

Designed by Nicholas Moll Design
Telephone: 020 8879 4080

Printed by The Ludo Press Ltd
Telephone: 020 8879 1881

CONTENTS

The Cardiac Surgery Demonstration Project Group

Leon Fine (Chairman)
University College London

Bruce Keogh (Co-chairman)
Queen Elizabeth Hospital, Birmingham
Society of Cardiothoracic Surgeons of Great Britain & Ireland

Mairi Gould (Project Coordinator)
University College London

Shan Cretin (Senior Scientist)
RAND, California

Maria Orlando (Associate Behavioural Scientist)
RAND, California

Participating surgeons and affiliated centres

Jon Anderson	Hammersmith Hospital, London
John Au	Victoria Hospital, Blackpool
John Hutter	Bristol Royal Infirmary, Bristol
Ben Bridgewater	Wythenshawe Hospital, Manchester
Jatin Desai	King's College Hospital, London
Leslie Hamilton	Freeman Hospital, Newcastle
Sam Nashef	Papworth Hospital, Cambridge
Charles Pattison	University College London Hospitals
James Roxburgh	Guy's & St Thomas' Hospital, London

Additional members of the Steering Group

Robert Brook	RAND, California
Nick Black	London School of Hygiene & Tropical Medicine
Jules Dussek	Past President, Society of Cardiothoracic Surgeons
John Wyn Owen	Nuffield Trust, London
Tom Treasure	Guy's & St Thomas' Hospital, London
Kathy Rowan	Intensive Care National Audit & Research Centre
Peter Walton	Dendrite Systems

FOREWORD

Research conducted over the past 25 years – mostly in the United States – has shown that quality of care is measurable, but dependent on the physicians, health care team and setting in which the patient is being treated; deficiencies cannot be ignored.

Recent events in the UK, and particularly, the Inquiry into Children's Heart Surgery at Bristol Royal Infirmary,[1] have emphasised the importance of recognising and acting on these findings. And there is no doubt that the Inquiry hearings changed public and professional opinion about the need for much greater public accountability than had been customary.

One of the key recommendations to emerge was the need for systematic internal and external monitoring of the clinical performance of health care professionals and hospitals, with the aim of identifying and sharing examples of good practice and flagging up poor performance.

The Society of Cardiothoracic Surgeons of Great Britain and Ireland (SCTS) long ago recognised the need for clinical monitoring to drive up standards and improve the quality of patient care through performance monitoring. It was the first speciality to address this, when it established the United Kingdom Cardiac Surgical Register in 1977.

The Register collects raw aggregated data from each cardiac unit on the outcomes of what now amounts to 40,000 surgical procedures carried out in the NHS every year.

Its annual reports failed to prevent the events at Bristol, because in the absence of risk adjustment, individual variance could be dismissed as a case mix issue. Inadequate data collection facilities and failure to track every death also skewed the figures for some procedures.

In a bid to encourage voluntary and honest compliance, the data were anonymised. But this actually served to cover up casual and inaccurate reporting, which had particular consequences in the complex arena of children's heart surgery, where categorisation of pre-operative risk can be extremely difficult.

Twenty years after the establishment of the Register, the government made it clear that the public had a right to know that professional standards were under scrutiny and that doctors could no longer hide behind anonymity. In the light of concerns raised by Bristol, the SCTS

understood the need for public reassurance and responded by requiring outcomes for individual surgeons to be included in the Register.

Confidential returns on individual surgeons' performance for "marker" operations on adult and children's heart surgery and thoracic surgery have been made to the Register since 1997. The results are subject to independent analysis and internal scrutiny by the SCTS, to detect any anomalies and ensure these are promptly addressed.

Clearly, when released into the public domain, data on units and individuals will provide an invaluable source of information to patients, surgeons, regulators and health care planners. But, crucially, their credibility hinges on the reliability and quality of the primary information – and the context in which they are delivered. In the case of individual performance data, most cardiothoracic surgeons remain convinced that risk adjustment is essential to take account of case mix.

The UK National Adult Cardiac Surgical Database, which the SCTS has run in tandem with the Register since 1994, collects detailed data for risk adjustment on procedures and patient outcomes for those centres participating in the scheme. However, ongoing concerns about the quality of these data have so far prohibited full and public disclosure of the three audit reports issued to date.

The Cardiac Surgery Demonstration Project was therefore designed to examine the robustness and credibility of data submitted to the Database, and to determine whether deficiencies could be rectified by external monitoring and validation to make it fit for issue into the public domain.

For its work on this project, the Nuffield Trust was fortunate to have been able to draw on the expertise and experience of RAND Health, with which it has been collaborating on the public release of health care information since 1999, under the aegis of the Nuffield Trust RAND Collaboration. This collaboration lies at the heart of the Trust's work on quality.

Achieving improvement in quality of care through publishing information about performance is an area in which the UK has, with notable exceptions, lagged behind the US. The largest non-governmental healthcare research organisation in the US, RAND Health specialises in the application of rigorous empirical research designs to health issues, including quality of care.

As well as producing a report on the evidence for the relationship between the provision of information about performance and quality improvement,[2] the Trust has funded several projects to explore the potential for the transatlantic transfer of the knowledge and techniques developed by RAND. The Cardiac Surgery Demonstration Project is the first of these.

The Trust has a long history of involvement in quality issues. In 1998 it published a report on UK health care quality initiatives by Professor Sheila Leatherman, who is now President of the Center for Health Care Policy at the University of North Carolina, and Senior Associate both of the Nuffield Trust and of the Judge Institute of Management at the University of Cambridge.

Her report, *Evolving Quality in the New NHS: Policy, Process and Pragmatic Considerations* contributed to the debate on achieving improvements in quality, started by the publication of the government's *The New NHS: Modern, Dependable* in 1997. Professor Leatherman is currently assessing progress at the mid point of the government's 10 year quality programme for the Trust.

Quality of information about health care is pivotal to assessing the extent to which quality initiatives actually improve quality of care, and how successful the government's programme has been. Without reliable, meticulous data, this will prove an impossible task.

The Cardiac Surgery Demonstration Project shows that it is possible to produce good quality, credible information, and, hopefully, it will inspire other specialities to take up the challenge. But the Project also emphasises the critical importance of risk adjustment, which should be universally adopted as the gold standard for any health outcomes data issued into the public domain.

John Wyn Owen CB
Secretary October 2003

EXECUTIVE SUMMARY

The hearings for the Bristol Royal Infirmary Inquiry changed public and professional opinion about the need for much greater public accountability of NHS health care organisations and those working within them than was customary.

One of the key recommendations of the Inquiry was that there should be systematic internal and external monitoring of the clinical performance of health care professionals and hospitals, with the aim of identifying and sharing examples of good practice and flagging up poor performance. The government responded, with a promise to publish high quality health information about the quality of care, relating to specific units and, in time, specific consultants.

Research shows that quality of care is measurable but dependent on the physicians, health care team or setting in which the patient is being treated. Quality of information about health care is pivotal to assessing the extent to which quality initiatives actually improve quality of care.

Cardiothoracic surgeons were the first specialist group to audit their own clinical results unit by unit and surgeon by surgeon, but the results had remained confidential. To prepare the results for public release, ensuring that the figures were fully adjusted to take account of disease severity, an assessment was made of the quality of data submitted to the National Adult Cardiac Surgery Database of the Society of Cardiothoracic Surgeons of Great Britain and Ireland.

The assessment entailed checking the reliability and completeness of information on first time isolated coronary artery bypass surgery, submitted to the Database over two periods totalling 17 months. Deficiencies and inaccuracies were fed back to the 10 units taking part in the study, with the aim of improving the data content and enhancing the accuracy of risk adjusted mortality figures.

The project revealed deficiencies in the completeness and reliability of the information supplied to the Database, with around 25% of essential information required for risk adjustment missing from the case records. Most of this was found in the case records after review, but had not been entered on to the Database.

Five months of monitoring, validation and feedback improved the overall quality, and significantly increased the amount of initial data provided. But participating centres found risk adjustment difficult and they were let down by poor quality NHS information technology, including the absence of computerised mortuary records.

But after risk adjustment had been recalculated, death rates across all the centres were very similar, providing public reassurance on the competence of NHS cardiothoracic units and emphasising the need to contextualise performance data in the light of case mix.

The poor response to a further request for data from the units involved after the project had been completed highlights the difficulties of sustaining a credible database once external surveillance has been withdrawn.

Without effective data capture, management and independent monitoring of data quality, it will be impossible to achieve the level of reliable information required for issue into the public domain, to improve clinical decision making and, ultimately, patient care.

To address this on a national scale for heart surgery will require a relatively modest investment to put in place the infrastructure required to improve and maintain the quality of information at the highest level. Up to now, the process has relied entirely on the goodwill and voluntary endeavour of a few members of the Society of Cardiothoracic Surgeons of Great Britain and Ireland, and this cannot be allowed to continue.

Recommendations:

- A comprehensive system-wide improvement in information technology networks for both administrative and clinical data is essential to provide reliable CABG data.
- A dedicated data manager should be employed in each heart surgery centre providing risk adjusted outcomes.
- A permanent cycle of independent external monitoring and evaluation of data quality should be established for a database providing information on cardiac surgery.
- The amount of data required must be manageable and appropriate to the available local resources, to prevent it from becoming an undue burden.
- A standardised system of risk adjustment/stratification for outcomes data on heart surgery should become part of clinical governance.
- Information on the performance of individual units and surgeons, intended for public release, should be validated by an independent source before issue.
- The potential problems relating to the long term use of confidential patient information, inherent in current legislation, must be addressed.
- The involvement of the speciality is essential to ensure high compliance with, and instil confidence in, public release of data used to judge performance, but to the extent that the public and the media will not lose faith in its independence.
- The inherent weaknesses of clinical outcome databases, as highlighted by the Cardiac Surgery Demonstration Project, must be fully recognised and communicated to the public in a comprehensible way, to avoid invidious and unhelpful comparisons between centres and individuals.

The Cardiac Surgery Demonstration Project

Aims

- To assess the quality and completeness of the existing National Adult Cardiac Surgery Database of the Society of Cardiothoracic Surgeons of Great Britain & Ireland (SCTS)
- To focus specifically on risk adjusted, inpatient mortality in isolated, first time coronary artery bypass graft (CABG) surgery across participating units
- To institute a programme of monitoring, validation and feedback with the aim of improving the quality of the information provided.

First time isolated CABG is the SCTS index procedure for adult heart surgery, and inpatient death, rather than death up to 30 days of surgery, was chosen as the outcomes indicator, because this was considered to be more clinically relevant and easier to validate. It is also used internationally by cardiothoracic surgical societies, and therefore readily bears comparison.

Ten tertiary care cardiac centres in England participated in the project. Each was assigned a centre number:

- Queen Elizabeth Medical Centre, Birmingham (1)
- Bristol Royal Infirmary, Bristol (2)
- Papworth Hospital, Cambridge (3)
- Victoria Hospital, Blackpool (4)
- Wythenshawe Hospital, Manchester (5)
- Freeman Hospital, Newcastle (6)
- Guy's & St Thomas' Hospital, London (7)

- King's College Hospital, London (8)
- Imperial College School of Medicine, Hammersmith Hospital, London (9)
- University College London Hospitals, London (10)

These centres had all contributed data to the SCTS database, and were selected on the basis of their geographic distribution and their perceived track records in data collection. In keeping with current policy, all data on individual patients and surgeons were anonymised, but the centres agreed that all the information generated should be made public for the purposes of this project.

The specific questions to be addressed were:

- Was there a substantial amount of missing data in the SCTS database, which would effectively invalidate risk adjusted mortality?
- Could the data elements entered into the database be validated by review of patients' records?
- Could recorded deaths be substantiated by the mortuary records?
- Could a period of monitoring, validation and feedback at participating centres improve the accuracy and completeness of the database?

A summary report of this project has already been published in the *British Medical Journal* (*BMJ* 2003; 326: 25–8).

A brief outline of the project

The project was carried out over two phases, the first of which was carried out retrospectively, and the second of which was carried out prospectively.

- For phase 1, case records for CABG previously submitted to the Database between April 1 1997 and March 1998 were grouped according to the degree of pre-operative risk (risk stratification).
- The project coordinator independently reviewed a random sample of 495 risk stratified case records to assess the completeness and reliability of the information originally provided by the centres, and recoded the risk scores, where necessary.
- A spot check was also made of the number of deaths declared by the centres for one month by cross checking these with mortuary records.
- The risk scores were then recalculated and compared with those originally submitted.
- For phase 2, the entire process was repeated at nine participating centres over five months on case records submitted from July 1 to November 30 2000.
- This time, the project coordinator reviewed 430 risk stratified case records, and recoded the scores, where appropriate.

- Scoring inaccuracies and incompleteness in the information supplied were regularly fed back to the participating centres throughout this period.
- After the project was completed, an attempt was made to cross-check the reported death rates against actual deaths for phase 2 figures.

Risk adjustment

Risk adjustment is required to rule out differences in outcomes attributable to disease severity, and to account for differing factors influencing the likelihood of dying as a result of the surgery at different centres.

Accurate risk adjustment is essential for the performance of a centre or an individual to be judged in context. In turn, the validity of risk adjustment relies on complete data collection for all risk factors, and for these to be consistently assessed and scored at all centres.

- Each patient is assessed before surgery (pre-operative risk) for a set of factors or variables, which, taken together, predict that patient's likely risk of death or some other well defined outcome.
- The selection of factors is based on previous studies of large numbers of patients where the exact contribution of each factor to a particular outcome can be determined statistically.

For example, for coronary bypass graft surgery, patients who are old, who are admitted as an emergency, who have poor heart function, or who have had previous heart operations, have a poorer chance of survival.

- Risk adjustment uses one of several methods to predict aggregate outcomes for a large group of patients to obtain overall predicted death rates for that centre.

For example, a centre serving patients with more severe disease and adverse risk factors will be more likely to register greater numbers of deaths as a result of surgery than a centre serving patients with less severe disease and fewer adverse risk factors.

- The predicted death rate for a centre is then compared with the numbers of actual or observed deaths.

Discrepancies between the actual and predicted death rates might simply be attributable to better or worse than predicted overall performance by that centre.

- Various statistical tests can be used to find out if the differences are of real clinical significance. Which test is used will depend on the risk adjustment method applied.
- In this study the standardised mortality ratio (SMR) was used. The SMR divides the actual death rate by the predicted death rate. If there is no difference between these, then the SMR equals 1.
- A 95% confidence interval is constructed around the SMR. This is a mathematical way of assigning the degree of confidence in the true level of risk. If the 95% confidence interval includes 1, the centre cannot be judged to be significantly better or worse than expected.

Risk score calculation

Table 1: Data elements collected. Those used to quantify pre-operative risk in the Parsonnet and EuroSCORE models are denoted as "required".

Variable	Parsonnet	EuroSCORE
ID - registry entry number		
Post code		
Date of birth (age)	Required	Required
Date of operation	Required	Required
Ethnic origin		
Sex	Required	Required
Angina - CCS[a] score		
Dyspnoea - NYHA[b] score		
Previous Q wave MIs[c]		
Last Q wave MI[c]		Required
Recently failed intervention	Required	Required
Diabetes	Required	
Renal system	Required	Required
Respiratory	Required	Required
Peripheral vascular disease		Required
Extent of coronary disease (inc. left main stem)		
Ejection fraction	Required	
Intra aortic balloon pump	Required	Required
Ventilated		Required
Operative priority		Required
First or redo surgery		
unstable angina (requiring intravenous nitrates)	Required	
Intravenous inotropes		Required
height	Required	
weight	Required	
hypertension	Required	
Operation performed by (grade)		
Cardiac procedures		
Patient status		

[a] Canadian Cardiac Society angina score [b] New York Heart Association dyspnoea score [c] Myocardial infarction

The Parsonnet score and the EuroSCORE (European System for Cardiac Operative Risk Evaluation) were used to score the risk of dying after surgery for each patient.

The Parsonnet score was the first simple, validated, additive scoring system used to predict risk in heart surgery (1989).[3] The EuroSCORE is a weighted additive score, similar to Parsonnet, but based on a pan-European sample of heart surgery patients. It was developed in 1999.[4]

To calculate the scores, 17 essential pieces of data were required for each patient. These were based on an initial list of pre-operative risk factors, drawn from several risk scoring systems: the UK Bayes model;[5] the US Bayes model;[5] the US CABG mortality reporting system;[6] the Parsonnet score and the EuroSCORE systems.

The 17 elements were **(table 1)**:

- Date of birth
- Date of operation
- Sex
- Last Q wave myocardial infarction or heart attack
- Recently failed intervention (necessitating surgery)
- Diabetes (history of)
- Renal system (severity of disease)
- Respiratory health (type)
- Peripheral vascular disease
- Ejection fraction (left ventricular function)
- Intra-aortic balloon pump (presence of before surgery)
- Ventilated (before surgery)
- Operative priority (from elective to salvage)
- Unstable angina requiring intravenous nitrates
- Use of intravenous inotropes
- Body mass index (height and weight)
- Hypertension or high blood pressure

How these elements were to be defined, assessed and scored was agreed in advance. This is critical: if different centres use different definitions or assessment methods, or routinely fail to collect some data elements, then the risk adjustment process produces biased or inaccurate results.

Based on that score, each patient was then assigned to a risk group (risk stratification). Risk group was the critical value rather than the crude score. There are five Parsonnet and six EuroSCORE risk groups **(table 2)**.

Table 2: Risk groups with corresponding EuroSCORE and Parsonnet score values

Risk group	1	2	3	4	5	6
Parsonnet score	0-4	5-9	10-14	15-19	>19	
EuroSCORE	0-1	2-3	4-5	6-7	8-9	>9

In the Parsonnet system an additive score of 0-4 equates to a pre-operative risk of death of 1% (low risk); 0-5 equates to 5% (increased risk); 10-14 equates to 9% (significantly increased risk); 15-19 equates to 17% (high risk); and a score of over 19 equates to 31% (very high risk).

For the EuroSCORE system, a score is assigned to specific factors and weighted. For example, every five years over the age of 60 scores 1 point, and being a woman scores 1

point, while neurological disease scores 2 points. A score of 0-1 is low risk, 2-3 is increased risk, and so on. A score of over 9 equates to very high risk.

Independent assessment

Just one coordinator reviewed the information supplied in the case records. This was to ensure that risk scores were recoded consistently, thereby enabling valid comparisons between centres to be made. The coordinator was not associated with any of the centres or the SCTS, so as to be able to act as an independent assessor.

As a training exercise, the coordinator independently reviewed three sets of anonymised case notes, which had already been coded by the lead surgeon, from each centre. Any discrepancies were resolved with the help of an experienced cardiac surgeon (BK). As a check on consistency, the coordinator recoded some of the records, without knowing where they were from, around two months later; in this validation exercise there was less than one discrepancy per patient record.

Retrospective evaluation of data submitted by each centre for the period 1 April 1997 to 31 March 1998

In all, the 10 centres submitted 7711 case records **(table 3)**. The formats varied considerably, from locally designed, handwritten forms to direct entry on to an electronic database. The software packages used also varied. Further details are provided in **Appendix: para 1**.

Any patient transferred out of the hospital and who died subsequently, or who died after being readmitted to the same hospital, was recorded as "alive" for the purposes of this study.

Risk adjusted death rates for each centre were calculated from the information and pre-operative risk scores supplied by each centre. The lack of data in some case records meant that some risk scores had to be estimated for these calculations.

Every case from every centre was scored in exactly the same way.

Table 3: Number of case records received in the two phases of the project (Phase 1: 12 month period; Phase 2: 5 month period)

	Phase 1		Phase 2	
Centre	Submitted	Re-abstracted	Submitted	Re-abstracted
1	896	55	247	50
2	826	53	382	48
3	1125	45	398	39
4	615	53	264	54
5	876	48	361	50
6	862	53	305	53
7	999	44	294	46
8	667	50	258	44
9	336	50	—	—
10	509	44	174	46
Total	7711	495	2683	430

Risk stratification

Risk stratification was carried out to ensure that a full spectrum of case severity was assessed. Around 54 records in each centre, amounting to nine cases in each of six risk categories, were randomly selected and stratified for severity, using the EuroSCORE. Some centres had a limited number of cases in the highest risk category, and consequently contributed fewer than nine cases.

Table 4: Criteria for validation of data elements

Data Element	Validation
1. Demographic	
Sex (M or F)	Addressograph
Date of Birth (DOB)	Administrative
Ethnicity	
2. Cardiac History	
CCS angina category (highest in 2 weeks preceding surgery)	pre-operative assessment and/or discharge summary
NYHA dyspnoea category	pre-operative assessment
Previous Q wave myocardial infarction	statement in pre-operative surgical / anaesthetic assessment
Last myocardial infarction	date noted from previous records /
*Euro*SCORE supplement (MI within 90 days)	pre-operative assessment
3. Previous non-surgical intervention	
Recently failed intervention	Doctors record (same admission)
Intravenous heparin / nitrates	Doctors record
Intravenous inotropes	Doctors record
4. Previous surgical intervention	
Previous CABG surgery	Doctors record
5. Risk factors for coronary disease	
Diabetes	Doctors record
Hypertension	Doctors record
6. Additional Medical History and Risk Factors	
Renal system	Doctors record and pre-operative serum creatinine level
Pulmonary system	Doctors record
Peripheral vascular disease (inc. carotid bruit)	Pre-operative assessment
7. Coronary Anatomy	
Extent of coronary vessel disease	Angiography report or clinician's statement
8. Ejection Fraction	
LV Function (EF)	Percentage from report or clinician's statement
9. Preoperative Support	
IV Nitrates or Heparin (include sc heparin)	Doctors record
Intravenous inotropes	Drug chart
Intra aortic balloon pump (IABP)	Doctors record
Ventilated	Doctors record
10. Operation Status	
Operative priority	Doctors record
Operation sequence	Doctors record
11. Bypass-related data	
Patient height	Perfusionist's / Anaesthetist's record or
Patient mass	pre-operative nursing assessment
12. Discharge or Death	
Patient status on discharge from hospital	Doctors record

This figure was based on the numbers of cases needed to detect meaningful differences between the submitted risk score and the recoded risk score **(Appendix: para 2)**. The coordinator then reviewed the stratified sample, without knowing the risk group assignment and searched for and recoded the 17 data elements on each of 495 case records.

Objective criteria for each risk factor agreed at the outset were used to validate the data **(table 4)**. The primary report – for example, the angiography report – was used wherever possible, but in many cases a test result or finding recorded in the doctor's notes was the validating source.

The submitted and recoded pieces of data for each centre were then compared. Risk adjusted death rates for each centre based on the originally submitted data were compared with those based on the risk stratified data.

Completeness and reliability of the database

The "percentage of missing data" for each centre was calculated as the average percentage of data elements missing across all patient records in that centre **(Appendix: para 3)**.

The kappa coefficient was used to assess reliability, because it measures the reliability of scoring by two sources and adjusts for agreement by chance alone **(Appendix: para 4)**.

The reliability score according to kappa is as follows:[7]

0.00	None
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 0.99	Almost perfect
1.0	Perfect

- Where possible, a kappa coefficient was calculated for each of the 17 data elements, to measure the extent of agreement between scores in the original data submitted by the centre and those recorded by the study coordinator after review.
- The "average kappa" for each centre was derived from the average of the kappa coefficients, calculated for each of the 17 data elements. But if any of the original or reviewed data elements are missing, or there are slight variations in the coding, the kappa statistic cannot be calculated.
- For the purposes of this study, an additional "centre reliability score" was computed by multiplying the average kappa for a centre by the proportion of total data elements for which kappa could be calculated **(Appendix: para 5)**.

In other words, if all 17 data elements are scored over the same range of values by both the centre and the coordinator, the centre reliability score would equal the average kappa. A wide discrepancy between a centre's average kappa and the centre reliability score indicates a large proportion of uncalculated kappas in that centre.

Mortuary record checks

Because the crude death rate for isolated CABG surgery is 2–3%, an error in the recording of only one or two deaths – either failure to record a death or incorrectly designating a death where the patient was alive at the time of discharge – would significantly influence the declared death rates.

Mortuary records in most UK institutions were not computerised at the time of this study; they were usually written by hand, increasing the chances of mistakes.

All submitted figures for actual deaths for one month declared by each centre were cross checked against mortuary records for that month and the ensuing 90 days – to take account of complications after surgery – to see if they matched. Names, ages or dates of birth were used to enhance the accuracy of the check. Hospital numbers, which are vital for identification, were not recorded in the mortuary records of four of the ten centres during that time **(table 5)**.

Table 5: Information held in mortuaries

Centre	Name	Age/DOB	Hospital No	Date of death
1	+	+	+	+
2	+	+	+	+
3	+	+	+	+
4	+	+		+
5	+	+		+
6	+	+		+
7	+	+	+	+
8	+	+	+	+
9	+	+		+
10	+	+	+	+

DOB = date of birth

Monitoring and feedback

Definitions of each data element were highlighted and illustrative cases presented at a meeting of study participants to maximise scoring accuracy. From the outset it was clear that the staff charged with data collection and entry varied widely. In some centres this was the responsibility of designated data managers; in others, the operating surgeon or a trainee surgeon carried this out.

Case record reviews for each centre were conducted on site for both phases of the study, during which the study coordinator flagged up problems, such as missing information and inconsistent or faulty scoring of data elements, with the lead surgeons and their teams.

The study coordinator visited each centre at least twice, and kept in regular e-mail and telephone contact to resolve misunderstandings or difficulties over the scoring criteria for certain data elements.

PHASE 2

Prospective analysis of data submitted by centres from 1 July to 30 November 2000

The process was repeated on 2683 patient records submitted from 1 July to 30 November 2000 for nine centres. A sample of 430 records was risk stratified, which were then reviewed and recoded by the study coordinator.

Mortuary data cross checks were also repeated, but only for five centres.

Findings

Issues arising from the validation process

Certain data items presented particular problems. In general, dates and times were not always clearly stated in case notes; often a day and month were present, but not the year. The filing of case notes was often very haphazard, and copies of referral information, operation notes and correspondence were often duplicated. The following data elements caused the most problems:

Identification
Because the data had been anonymised, extra work and time was needed to track the hospital number and check the notes.

Cardiac history: angina status (Canadian Cardiovascular Society) and Dyspnoea (New York Heart Association) Scores
There was often no written description of the patient's symptoms preceding surgery and therefore only the assigned score on the clerking chart could be located. Where symptoms had been recorded, often only chest pain or shortness of breath had been noted. These two items were modified before phase 2 so that they were assessed as the most severe on admission or within two weeks before surgery.

Previous Q wave myocardial infarction
The type of myocardial infarction was not always stated.

Last Q wave myocardial infarction
Dates and times were not always clearly stated.

Coronary anatomy
Angiography reports were not always present; "4-vessel" disease was often described.

Height
This information was frequently missing and often recorded in imperial measures, which may account for some of the discrepancies in the existing databases.

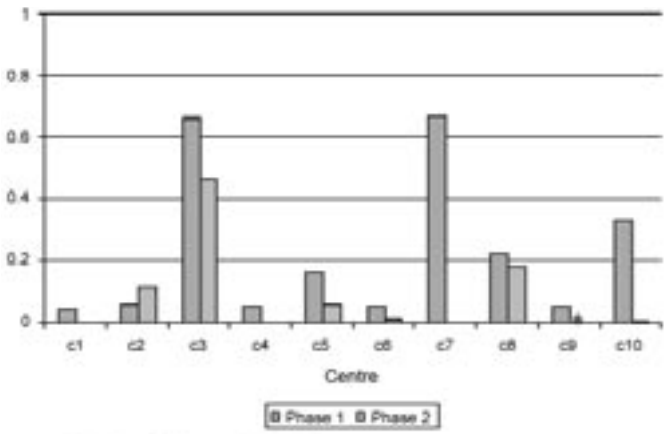
Weight
This was often recorded in imperial measures, which may account for some of the discrepancies in the existing databases.

These problems were emphasised during feedback.

Completeness and reliability of the database

Completeness of the database
During phase 1, 10 centres submitted sets of data but centre 9 submitted data on only 40% of their cases. During phase 2, technical difficulties prevented centre 9 from capturing any data electronically – one of the conditions of participation in the study.

Fig 1: Proportion of missing data on data submitted by Centres for stratified sample of case records.†

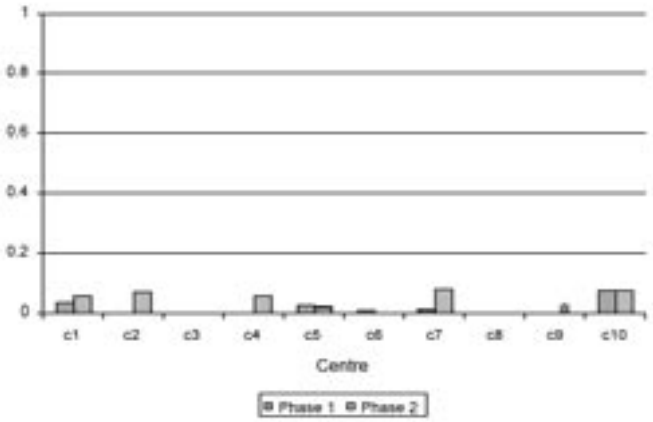


* No Phase 2 data submitted
† Based upon 17 data elements required to calculate EuroSCORE and Parsonnet Score .

Figure 1 shows the completeness of submitted data for phases 1 and 2 for individual centres. **Figure 2** summarises the completeness of the reviewed data for phases 1 and 2.

The proportion of missing data in the reviewed case notes was minimal in both phases: 1.81% (SE 0.06%) in phase 1 and 4.04% (SE 0.05%) in phase 2. This suggests that in centres with high proportions of missing data, the information was in the records but had not been transferred on to the database.

Fig 2: Proportion of missing data for stratified sample of case records when re-abstracted by study co-ordinator. †



* No Phase 2 data submitted
† Based upon 17 data elements required to calculate EuroSCORE and Parsonnet Score.

The percentage of missing data between the submitted records in phase 1 (24.96%; SE 0.09%) and those submitted in phase 2 (9.33%; SE 0.08%) fell significantly ($p<0.001$) **(table 6)**.

Table 6: Outcomes of the project: completeness and reliability of the database

	Phase 1			Phase 2		
	% Missing Data Elements (±SE)	Average Kappa (±SE)	Centre Reliability Score (±SE)	% Missing Data Elements (±SE)	Average Kappa (±SE)	Centre Reliability Score (±SE)
Submitted	24.96 (0.09)	0.67 (0.11)	0.44 (0.17)	9.33 (0.08)	0.78 (0.06)	0.53 (0.15)
Reabstracted	1.81 (0.06)			4.04 (0.05)		

(±SE) = standard error

Average kappa refers to comparisons between submitted and reabstracted data across all centres.
See text for definition of Centre Reliability Score.

Reliability of the database

In all but one centre the average kappa value increased between phases 1 and 2, reflecting an overall improvement in centre reliability score **(figure 3)**.

During phase 1, the average kappa across nine centres for submitted and reviewed data elements was 0.67 (SE 0.11), indicating a "substantial" level of reliability. But the large amount of missing data and data coded over different ranges of values, meant that only relatively few kappa values were calculated in each centre **(table 7)**. The resulting average centre reliability score of 0.44 (SE 0.17) therefore reflected only a "moderate" level of reliability.

Fig 3: Average Kappa coefficients for submitted versus re-abstracted data elements for the two phases of the project.

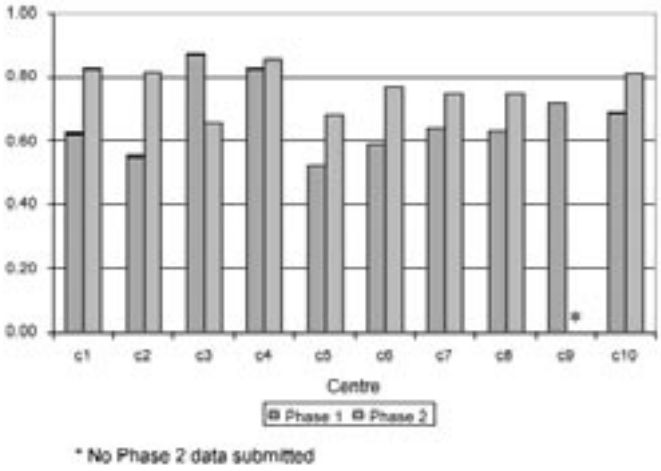


Table 7: Number of kappas calculated in each centre out of 17 data elements by Phase.

Calculated kappas		
Centre	**Phase 1**	**Phase 2**
1	16	14
2	12	10
3	4	6
4	16	14
5	11	8
6	15	12
7	5	12
8	11	10
10	12	15

During phase 2, the reliability of the information in both the submitted and reviewed sets of data improved, although not significantly (p 0.189), producing an average kappa of 0.78 (SE 0.06).

But the average centre reliability score, adjusted for the proportion of kappa coefficients calculated for each centre, was 0.53 (SE 0.15), indicating an overall improvement after the period of monitoring. But this was not significant (p=0.345), and the final centre reliability score was only "moderate".

Data for individual centres for phases 1 and 2 are shown in **figure 4**.

Calculation of risk scores

Ten centres submitted data for phase 1 and nine for phase 2 to enable the EuroSCORE to be calculated. For the Parsonnet score, eight centres submitted sufficient data in phase 1 and seven in phase 2.

Fig 4: Average Centre reliability scores for submitted versus re-abstracted data elements for the two phases of the project.

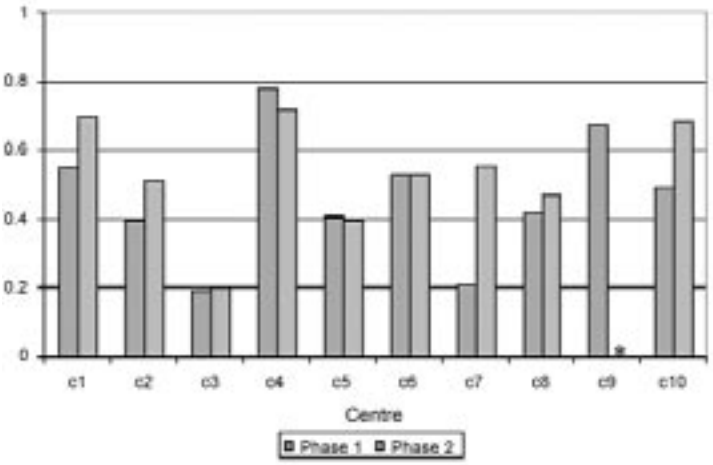


Fig 5: Kappa coefficients for Parsonnet risk groups, reflecting the level of agreement between submitted and re-abstracted data.

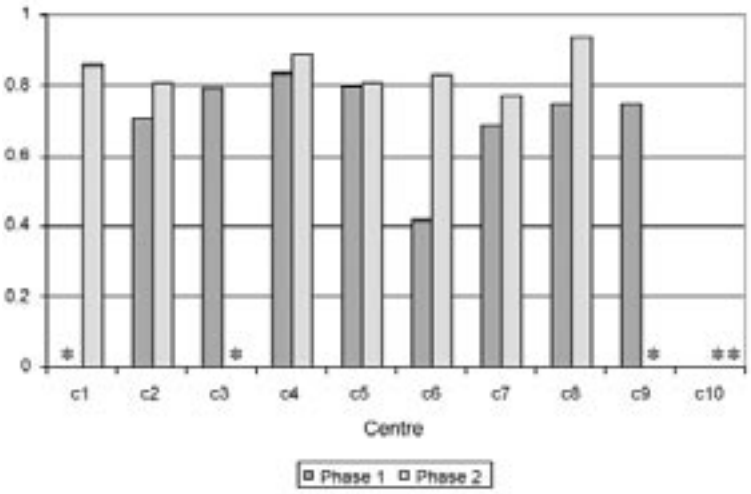
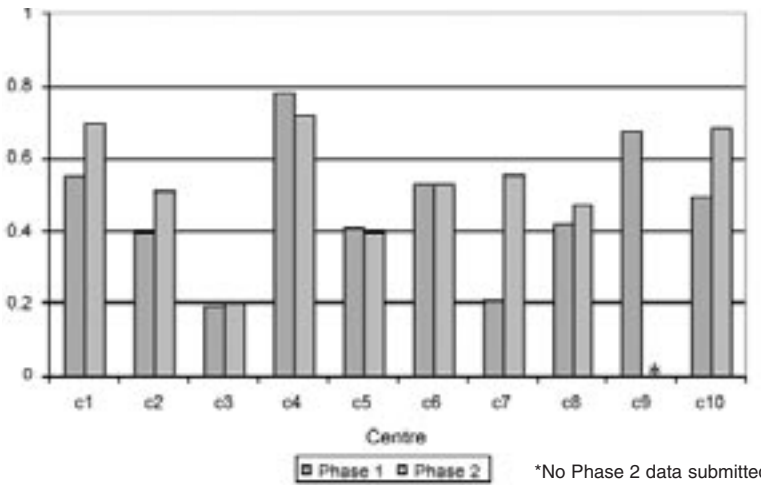


Fig 6: Kappa coefficients for EuroSCORE risk groups reflecting the level of agreement between submitted and re-abstracted data.



*No Phase 2 data submitted

The degree of agreement in risk group assignment for submitted and reviewed data improved from a kappa of 0.70 (phase 1) to 0.84 (phase 2), using the Parsonnet score **(figure 5)**. But persistent problems in coding some of the data elements for the EuroSCORE meant that the kappa value hardly changed: 0.65 (phase 1) and 0.61 (phase 2) **(figure 6)**.

Predictive value of risk adjusted mortality

Actual death rates minus predicted death rates were calculated for each centre, using both the EuroSCORE and Parsonnet systems. And the risk adjusted death rates were compared across the centres.

Using the entire set of submitted data – 7711 records in phase 1 and 2579 in phase 2 – predicted mortality was calculated by applying overall number of deaths in each risk group to the distribution of EuroSCORE and Parsonnet scores in each centre.

Table 8: Actual minus risk adjusted mortality correlation coefficients and rank order coefficients comparing Parsonnet and EuroSCORE risk-adjusted mortality in Phase 1 and Phase 2

	Phase 1		Phase 2	
	Actual - Predicted Mortality	Rank Order	Actual - Predicted Mortality	Rank Order
Correlation	0.922	0.915	0.951	0.933
Number of centres	(n=10)		(n=9)	

Using either the EuroSCORE or the Parsonnet score, the risk adjusted death rates were similar across all the centres **(table 8)**. The correlations were marginally higher after monitoring and feedback, but the differences were not significant.

Validation of deaths using mortuary data

Discrepancies in the numbers of deaths recorded by each centre and mortuary records were noted in three out of 10 centres during the phase 1 check **(table 9)**.

The declared death rate for centre 2 (Bristol Royal Infirmary) was 0.61% for 1997–8. But for the checked month, three out of 67 patients died, giving a death rate of 4.48%. This was attributable to problems in data transfer from one software system to another, and not reporting inaccuracies.

Table 9: Mortuary check for Phase 1

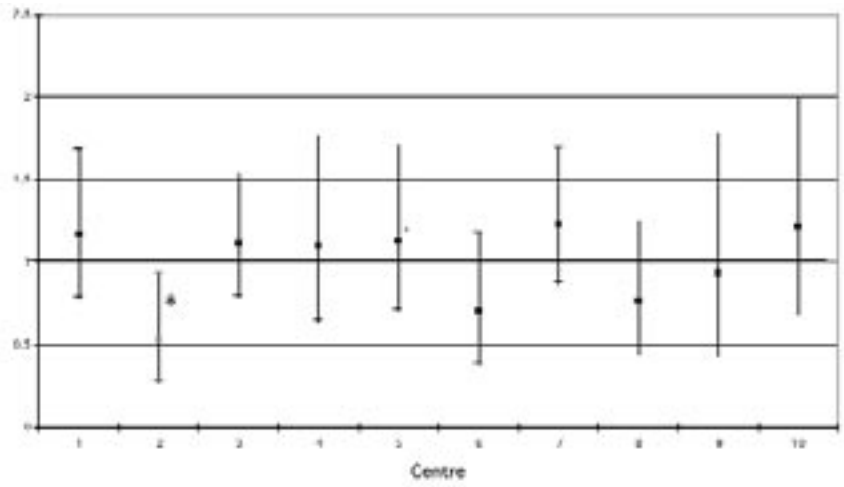
Centre	Check month	Mortuary record of deaths/cases	Mortuary check
1	June	4/71 (5.63%)	no discrepancies
2	May	3/67 (4.48%)	discrepancies
3	November	1/86 (1.16%)	no discrepancies
4	May	1/44 (2.27%)	no discrepancies
5	May	1/73 (1.37%)	no discrepancies
6	May	2/59 (3.39%)	discrepancies
7	May	6/88 (6.82%)	discrepancies
8	May	1/45 (2.22%)	no discrepancies
9	May	1/49 (2.04%)	no discrepancies
10	July	2/52 (3.85%)	no discrepancies

In phase 2, mortuary checks were carried out at five centres. Once again, there were discrepancies in two: Freeman Hospital (centre 6) had not entered all deaths on the cardiac database; and there was a mismatch with the handwritten hospital number recorded in the mortuary at Guy's and St Thomas' Hospital (centre 7) between database entry and mortuary record.

Actual and risk adjusted mortality (SMR)

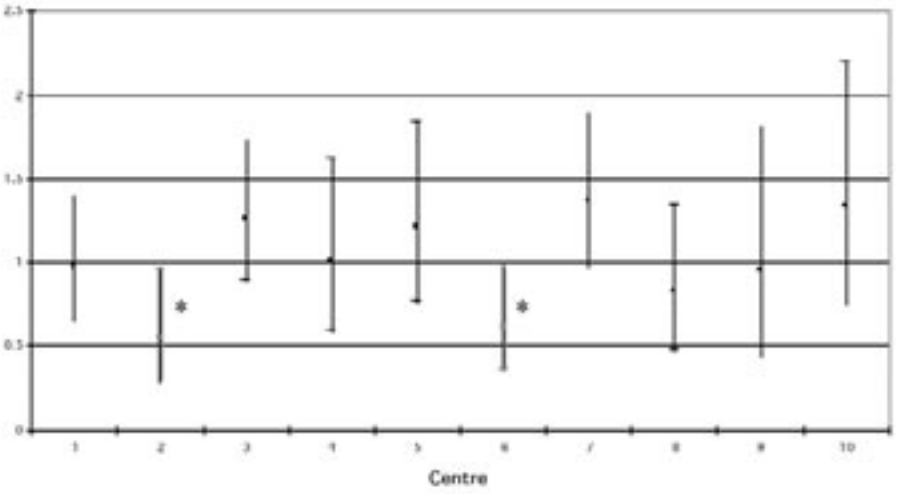
The standardised mortality ratios (SMRs) and 95% confidence intervals for each centre in phase 1, are shown for the Parsonnet (**figure 7**) and EuroSCORE (**figure 8**) risk scoring systems.

Fig 7: Parsonnet-adjusted standardised mortality ratios for participating Centres in Phase 1



* significantly different from 1

Fig 8: EuroSCORE-adjusted standardised mortality ratios for participating Centres in.



* significantly different from 1

During phase 1, the actual death rate was significantly better than expected at Bristol Royal Infirmary (centre 2), using either EuroSCORE or Parsonnet risk adjustment. At Freeman Hospital (centre 6), the actual death rate was better than expected, using both risk adjustment systems, but only seemed to be significantly better than expected, using EuroSCORE.

The discrepancies between mortuary records and submitted death rates for both of these centres might explain this.

Declared death rate check

The risk adjusted death rates may appear better than expected if a centre fails to record all deaths on the database. And one or two centres seemed to have an actual death rate that was substantially lower than expected after risk adjustment.

Given the discrepancies between the death rates declared by centres and the mortuary records, the figures for phase 2 were checked again after the project had finished and before phase 2 risk adjusted mortalities for the centres were compared.

Four months after the request had been made only five centres resubmitted the data from phase 2. In two centres the originally submitted numbers were confirmed. But in three centres small differences emerged, and the originally submitted figures were lower by 2/412, 2/333 and 5/287 cases, respectively. These errors substantially affect risk adjustment of death rates.

Uncertainty about these statistics in the other centres precluded risk adjusted death rates from being included for phase 2.

Key findings

- This sample of 10 centres revealed deficiencies in the completeness and reliability of the information in the SCTS National Adult Cardiac Surgery Database.
- Five months of monitoring, validation and feedback improved the completeness and reliability of the information submitted to the Database, but there was substantial room for further improvement, and this was after the centres had been involved in the project for almost two years.
- Around 25% of the essential information required for risk adjustment had not been entered onto the Database; this fell to 9% after monitoring and feedback.
- Most of the missing data required for risk adjustment were present in the case notes but had not been entered into the database.
- Participating centres found risk adjustment difficult to complete and the results were subject to inaccuracy.
- After recalculation of risk adjustment, death rates across all 10 centres were very similar (within two standard deviations).
- Mortuary records were handwritten, prone to error, and lacked basic data, such as hospital numbers, which compromised the accuracy of declared mortality figures.
- Many different grades of staff were involved in the handling of data collection and entry.
- Inadequate information technology systems within the NHS compromised data handling.
- The poor response to a request for data submission after the project had completed highlights the difficulties of sustaining a credible database once external surveillance has been withdrawn.

Specific issues arising from this project

Clearly, it is difficult to achieve maximum reliability for a database even when an external monitoring system is in place, but there are several possible explanations for this.

The level and frequency of contact between the study coordinator and individuals in the centres involved in data handling might have been less than what was required. This is a logistics issue, with time and cost implications.

The level of professionalism of those charged with data management at the centres might not have been optimal. Data were entered variously by operating surgeons, surgical trainees, secretaries, or a designated data manager. Funding for dedicated cardiac data managers has not been a priority for many NHS trusts, and staff turnover in other areas will inevitably compromise further data handling.

Two centres emerged with substantially better than expected death rates at the end of phase 2. This could have been due to genuine levels of excellence, but, equally, it could have been due to incorrect registering of the number of actual deaths.

A further check on five centres showed that three had entered the number of deaths incorrectly. This obviously affected their mortality statistics relative to other centres and precluded the publication of the final figures for death rates for phase 2.

These problems are in part attributable to the extremely heavy workload of clinical and administrative staff. But the part played by the wholly inadequate information technology provision in the NHS cannot be ignored. A case in point is the hopelessly inadequate mortuary record system, which made it difficult even to access hospital deaths in an electronic format for most hospitals.

Broader implications

The project shows that the SCTS has developed a sound basis for addressing quality issues, even if, as might be expected, there is room for improvement. And it shows that the UK is capable of producing information about performance that not only mirrors the quality of what is available in the US, but is also credible and useful to purchasers, providers and users of health care services alike.

Of prime importance, however, is the understanding that the process of maintaining a high quality database is continuous. Only five out of the nine participating centres responded to a request to resubmit their mortality data after the project had completed. This was despite reminders and a four month grace period. Even centres that have been involved in a data collection exercise fall down in their willingness and ability to sustain a credible database once external surveillance has been withdrawn.

The notion that local monitoring, through a process of clinical audit, should be at the core of performance for all trusts is a key tenet of NHS modernisation. And possibly the most effective strategy would be to use external monitors to scrutinise and validate cardiac surgical data and provide feedback to those responsible for the process locally.

The advantage of this would be that the monitors could carry out the same exercise at several trusts and could share experiences of best practice and common misconceptions and errors. But this presupposes that the monitor is independent of any trust and would not therefore be under pressure to issue reports based on incomplete or inaccurate information.

The complete independence of the project chairman and the study coordinator from the participating centres and from the SCTS was a major strength of this project. But to extend this approach to all cardiac centres in Great Britain and Ireland, would require the creation of a small, independent organisation, with a small number of coordinators, and an independent analytical and statistical arm.

This begs the question of who should fund the process.

It is abundantly clear that without effective data capture management and independent monitoring of data quality, it will be impossible to achieve the level of reliable information required to improve clinical decision making and, ultimately, patient care.

And it will be impossible to guarantee the uniformity of data validation and presentation, to permit meaningful comparisons between hospitals and between doctors. Until such a

system is in place, the future release of outcomes data on health care information will be vulnerable to professional and public doubts about its validity.

To address this on a national scale will require relatively moderate investment to put in place the infrastructure required to improve and maintain the quality of information at the highest level.

None of this is to minimise professional pride of ownership of a local database. Through professional societies such as the SCTS, surgeons can compare their outcomes, with the common goal of minimising the variances, rather than creating potentially damaging league tables, subject to poor risk adjustment and poorly validated data.

Dr Foster, the independent company which publishes comparative tables on UK health outcomes, based its consumer guide on the Department of Health's Hospital Episode Statistics (HES). Despite the moves to improve the accuracy and scope of HES, so that they can be used to underpin the clinical governance framework and flag up possible areas of concern, HES were not designed to collect detailed quality clinical data.

And simply ranking units or individual surgeons by the numbers of patients dying after surgery is not necessarily the most effective way of assuring safety or creating better informed patient choice. The sickest patients are the ones most likely to benefit from surgery, but they also run the greatest risk of dying as a result.

Significantly, this project showed that after risk adjustment, all participating centres had very similar death rates. The figures indicate that patients can be reassured about the competence of cardiac surgical services throughout the NHS. But they also emphasise the vital importance of risk adjustment to put data as sensitive as these into their proper context. This is particularly relevant to outcomes data for individual surgeons.

When it comes to the measurement of death rates, clearly, UK services have nothing to fear from a comparison of quality with the US. And international benchmarking of this kind has a major contribution to make to quality assurance for the benefit of both service providers and patients.

Around 97 per cent of patients survive coronary artery bypass surgery in both the UK and the US. For a highly technical procedure involving cardiologists, heart surgeons, nurses and technicians, these figures bear testimony to the achievements of effective teamwork. A zero death rate is unachievable, but year on year, the survival results have continued to improve, despite the fact that patients are increasingly older and sicker. But inevitably, there will be differences between hospitals.

The future of data quality

Following the recommendations of the Bristol Inquiry, and the government's response to it in 2002, the notion of professional competence has undoubtedly broadened. More precise definitions of standards of care and performance monitoring are becoming integral to NHS culture.

A data quality indicator is included in the annual NHS Performance Indicators, and the Commission for Health Improvement (CHI) looks at the quality of data available to trusts in its regular clinical governance reviews.

But overall progress has been somewhat patchy. The data quality strategy to support the NHS Modernisation Agenda, which was to have been developed by the NHS Information Authority by September 2002, has had to be rethought.

According to the National Institute for Clinical Excellence,[8] the strategy aimed to embrace all roles and levels of responsibility, include training needs, and make data quality an integral part of all data sets, with feedback as a key driver to data quality. The strategy has now devolved to CHI and the Department of Health Information Policy Unit.

A complementary focus of the Coronary Heart Disease Information Strategy, which has been designed to support the National Service Framework, is the creation of an information structure. This is to be achieved through data sets and clinical audit database development "to allow access to consistent and comparable information for clinical governance, performance monitoring, service planning, and public health."[9]

In addition to the SCTS Register and the National Cardiac Surgical Database, the UK Heart Valve Registry, has collected survival data on about 60% of all patients undergoing heart valve surgery since 1986. Now linked to the Office of National Statistics (ONS), its success spurred the Department of Health to proceed with the Central Cardiac Audit Database (CCAD), to which all 30 cardiac units in England will be electronically networked.

The CCAD, which is also linked to the ONS, aims to track a patient's treatment progress, irrespective of the hospital at which it is carried out, until death. A pilot project of six centres, which reported in 2000, highlighted serious technical problems.

While most of these seem to have been ironed out, progress linking units to the CCAD has been extremely slow, and there have been some difficulties merging the data for the National Cardiac Surgical Database. As a result the network will not be operational before the scheduled publication of the 30 day mortality rates of every heart surgeon in England by April 2004.

Data protection is also a potentially thorny issue. The Data Protection Act 1998 governs, in broad terms, the processing of information relating to living individuals. Clause 60 of the Health and Social Care Act 2001 provides additional and more specific restrictions on the use of information relating to patients.

This legislation still enables certain prescribed services to continue using confidential patient information without explicit consent. Simple benchmarking and outcome measures can, of course, be generated using anonymised patient data, but patient identifiers to facilitate more sophisticated long term follow up may, in time, require patient consent. Many registry coordinators fear that this will compromise complete data collection on an unpredictable scale.

Conclusions

The Society of Cardiothoracic Surgeons has shown what can be achieved. The next steps will be to extend this approach on a permanent and UK wide basis, and to explore whether this project can provide a template for other specialties to adapt for their own use.

But it cannot be emphasised too strongly how carefully performance data should be checked before they become public property. The New York State Department of Health spent almost three years validating its 1998 data before publication. The Department of Health recognised the need for robust data in its 2002 response to the Kennedy Report of the Bristol Inquiry.[11]

> "Our aim is to publish high quality health information about the quality of care, which can be related to specific units and, in time, specific consultants … It will take time as well as commitment to make sure the data are robust, that the clinicians are confident that it tells the whole story, and that it informs, rather than confuses, the user."

It will take only a few instances of the discovery of erroneous or inaccurate information to undermine confidence in the whole system. The Cardiac Surgery Demonstration Project highlights just how fallible and brittle the system really is.

Recommendations

- A comprehensive system-wide improvement in information technology networks for both administrative and clinical data is essential to provide reliable CABG data.
- A dedicated data manager should be employed in each heart surgery centre providing risk adjusted outcomes.
- A permanent cycle of independent external monitoring and evaluation of data quality should be established for a database providing information on cardiac surgery.
- The amount of data required must be manageable and appropriate to the available local resources, to prevent it from becoming an undue burden.
- A standardised system of risk adjustment/stratification for outcomes data on heart surgery should become part of clinical governance.
- Information on the performance of individual units and surgeons, intended for public release, should be validated by an independent source before issue.
- The potential problems relating to the long term use of confidential patient information, inherent in current legislation, must be addressed.
- The involvement of the speciality is essential to ensure high compliance with, and instil confidence in, public release of data used to judge performance, but to the extent that the public and the media will not lose faith in its independence.
- The inherent weaknesses of clinical outcome databases, as highlighted by the Cardiac Surgery Demonstration Project, must be fully recognised and communicated to the public in a comprehensible way, to avoid invidious and unhelpful comparisons between centres and individuals.

References

1 *The Report of the Public Inquiry into Children's Heart Surgery at Bristol Royal Infirmary 1984-1995*. The Bristol Royal Infirmary Inquiry, 2001.

2 Marshall M, Shekelle P, Brook R, Leatherman S. Dying to know. *Public release of information about quality of care*. London: Nuffield Trust-RAND Collaboration, 2000.

3 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 (suppl 1)* 1989; 6: 3-12.

4 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. *European Journal of Cardiothoracic Surgery* 1999; 15: 816-23.

5 Keogh B, Kinsman R. *National Adult Cardiac Surgery Database Report 1999-2000*. London: Society of Cardiothoracic Surgeons of Great Britain and Ireland 2001.

6 Jones RH, Hannan EL, Hammermeister KE, Delong ER, O'Connor GT, Luekper RV, *et al*. Identification of pre-operative variables needed for risk adjustment of short term mortality after coronary artery bypass surgery. The Working Group Panel on the Cooperative CABG Data Project. *Journal of the American College of Cardiology* 1996; 28: 1478-87.

7 Landis J R, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1997; 33: 159-74.

8 National Institute for Clinical Excellence. *Principles for Best Practice in Clinical Audit*. 2002; Oxford, Radcliffe Medical Press.

9 Information Policy Unit, Department of Health. National Clinical Audit Support Programme: Coronary Heart Disease Information Strategy. www.doh.gov.uk/ipu/ncasp/ncaspadv2.htm

10 Carlisle D. Change of heart. *Health Service Journal* May 15 2003: pp 12-13.

11 Department of Health. *Learning from Bristol: The Department of Health Response to The Report of the Public Inquiry into Children's Heart Surgery at Bristol Royal Infirmary 1984-1995*. London: TSO, 2002.

Appendix

1 Data collection

The anonymised data from each centre were submitted in Excel format and standardised by the study coordinator. Data analyses were carried out using Microsoft Excel and the Statistical Package for the Social Sciences (SPSS). Parsonnet scores and EuroSCOREs were calculated using fixed algorithms so that each case from each centre was scored uniformly.

2 Power calculation to determine sample size

The means and standard deviations for the EuroSCORE from each centre were used to calculate the sample size necessary to detect a meaningful difference between the submitted risk score and the recoded risk score. Around 50 sets of case notes for each centre were required to achieve 80 per cent power to detect a mean difference of one standard deviation in EuroSCORE between the original and recoded records.

3 Proportion of missing data

The proportion of missing data was calculated for each of the 17 data elements for the sampled records submitted by each centre and then reviewed by the study coordinator.

For example, in phase 1, there were 45 patient observations each in the submitted and reviewed datasets for centre 3. The data element "Ejection Fraction" was missing for five of the 45 observations in the submitted data (0.111), and for one of the 45 observations in the reviewed data (0.022). Proportions missing were calculated for each of the 17 data elements in each of the two datasets for each centre in each phase (a total of 17 x 2 x 9 x 2=612 proportions).

For each phase of the study, two centre level measures of missing data were computed, by averaging across all 17 required data elements to construct: (1) the proportion of missing data in the files submitted by each centre; and (2) the proportion of missing data when the same patient records were reviewed by the study coordinator. Finally, the overall proportion of missing data submitted and reviewed for each phase was estimated by computing averages across the nine centres.

Standard errors for the submitted and reviewed proportion of missing data in each phase were calculated from the variance of each centre and the average variance across centres.

A centre specific variance was calculated as the sum of the column totals of each centre's 17 x 17 variance-covariance matrix of the proportions of missing data elements, divided by $(17 \times n)^2$, where n is the number of observations for the centre. The average variance was estimated as the sum of the centre variances divided by 81 (number of centres squared).

Finally, the standard error across centres was taken as the square root of the average variance multiplied by 100 (to put it in the percentage metric). Z-tests were used to test significant differences between proportions of missing data values.

4 Kappa coefficients

The standard errors for the overall kappas were calculated by bootstrapping the nine centre specific kappas, using 2000 replications. The significance of the difference between the overall kappas from phases 1 and 2 was evaluated using the z-test ($z=0.878$, $p=0.189$).

Kappa coefficients could not be calculated if either submitted or reviewed data were missing. The kappa coefficient is based on frequency tables and can only be calculated for square tables. Therefore, if a centre coded a data element using a scale from 0-3 and the study coordinator correctly used a scale from 1-4, or the submitted data originally contained only zeros, while the reviewed data contained one observation with a value of 1, the kappa coefficient could not be calculated. Similarly, this applied if a centre submitted data with values of 0, 1, 2, and 3 while the reviewed data contained values of 0, 1, and 2, despite the fact that the two sets of data might have a great extent of agreement.

5 Centre reliability scores (adjusted kappa)

As with the overall kappas, overall reliability scores across centres for each study phase were the averages of the individual centre reliability scores. The standard errors for the overall reliability scores were calculated by bootstrapping the nine reliability sores from each of the centres, using 2000 replications. The significance of the difference between the overall reliability scores from phases 1 and 2 was evaluated using the z-test ($z=0.397$, $p=0.345$).