

- 24 Zethraeus N, Johannesson M, Jonsson B. A computer model to analyze the cost-effectiveness of hormone replacement therapy. *Int J Technol Assess Health Care* 1999;15:352-65.
- 25 Col NF, Eckman MH, Karas RH, Pauker SG, Goldberg RJ, Ross EM, et al. Patient-specific decisions about hormone replacement therapy in postmenopausal women. *JAMA* 1997;277:1140-7.
- 26 Col NF, Pauker SG, Goldberg RJ, Eckman MH, Orr RK, Ross EM, et al. Individualizing therapy to prevent long-term consequences of estrogen deficiency in postmenopausal women. *Arch Intern Med* 1999;159:1458-66.
- 27 Hillner BE, Hollenberg JP, Pauker SG. Postmenopausal estrogens in prevention of osteoporosis. Benefit virtually without risk if cardiovascular effects are considered. *Am J Med* 1986;80:1115-27.
- 28 Rymer J, Wilson R, Ballard K. Making decisions about hormone replacement therapy. *BMJ* 2003;326:322-6.
- 29 Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med* 1977;296:716-21.
- 30 Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *BMJ* 1995;311:1356-9.
- 31 Daly E, Gray A, Barlow D, McPherson K, Roche M, Vessey M. Measuring the impact of menopausal symptoms on quality of life. *BMJ* 1993;307:836-40.
- 32 Easton DF, Ford D, Bishop DT. Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Am J Hum Genet* 1995;56:265-71.
- 33 Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* 1998;62:676-89.
- 34 MacLennan A, Lester S, Moore V. Oral oestrogen replacement therapy versus placebo for hot flushes (Cochrane Review). In: *Cochrane Library*. Issue 1. Chichester: John Wiley, 2004.
- 35 Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst* 1998;90:1371-88.
- 36 Zethraeus N, Johannesson M, Henriksson P, Strand RT. The impact of hormone replacement therapy on quality of life and willingness to pay. *Br J Obstet Gynaecol* 1997;104:1191-5.
- 37 Hays J, Ockene JK, Brunner RL, Kotchen JM, Manson JE, Patterson RE, et al. Effects of estrogen plus progestin on health-related quality of life. *N Engl J Med* 2003;348:1839-54.
- 38 Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;2:iv,1-74.
- 39 Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998;351:123-7.
- 40 Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. Collaborative Group on Hormonal Factors in Breast Cancer. *Lancet* 1997;350:1047-59.
- 41 Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Med Decis Making* 1993;13:322-38.
- 42 Medicines and Healthcare products Regulatory Agency. Use of hormone replacement therapy in the prevention of osteoporosis: important new information. www.mhra.gov.uk (accessed 19 Dec 2003).

(Accepted 6 January 2004)

Retrospective cohort study of false alarm rates associated with a series of heart operations: the case for hospital mortality monitoring groups

Jan Poloniecki, Charalambos Sismanidis, Martin Bland, Paul Jones

Abstract

Objective To examine the efficacy of different methods of detecting a high death rate and determining whether an increase in deaths after heart transplantation could be explained by chance.

Design Retrospective analysis of deaths after heart transplantation. Seven methods were used: mortality above national average, mortality excessively above national average, test of moving average mortality, test of number of consecutive deaths, sequential probability ratio test (SPRT), cusum with v-mask, and CRAM chart. The national average mortality was not available, and a rate of 15% was used instead as the benchmark.

Setting Regional cardiothoracic unit.

Participants All 371 patients who received a heart transplant in the programme, 1986-2000.

Main outcome measures 30 day survival after transplantation.

Results All methods provided evidence that the 30 day mortality had been high at some stage. The probability that the finding was a false positive depended on which test was used. At the end of the series the average mortality, sequential probability ratio, and cusum tests indicated a level of deaths higher than the benchmark while the remaining four tests yielded negative results.

Conclusions If the decision to test for outlying mortality is made retrospectively, in the light of the data, it is not possible to determine the false positive rate. Prospective on-site mortality monitoring with the

CRAM chart is recommended as this method can quantify the death rate and identify periods when an audit of cases is indicated, even when data from other hospitals are not available. A hospital mortality monitoring group can routinely monitor all deaths in the hospital, by specialty, using hospital episode statistics (HES) data and appropriate statistical methods.

Introduction

In September 2000 heart transplantation at St George's Hospital, London, was suspended because of concern that more patients were dying than previously. The newspapers reported that 80% mortality in the last 10 cases had been of particular concern because this was "more than five times the national average."¹ We tested these assumptions—that surgical results had been satisfactory but later became unsatisfactory—against numerical criteria.

Methods

We examined seven tests that were available for comparing deaths with a benchmark death rate. None

Editorial by de Leval and p 379

Community Health Sciences, St George's Hospital Medical School, London SW17 0RE

Jan Poloniecki
senior lecturer
Charalambos Sismanidis
research assistant
Martin Bland
professor

St George's Healthcare NHS Trust, London SW17 0QT
Paul Jones
medical director

Correspondence to:
J Poloniecki
j.poloniecki@sghms.ac.uk

BMJ 2004;328:375-9



Detailed statistical methods of determining the false positive rate and an extra table of data can be found on bmj.com



This is the abridged version of an article that was posted on bmj.com on 29 January 2004: <http://bmj.com/cgi/doi/10.1136/bmj.37956.520567.44>

Table 1 Number of transplant operations and deaths within 30 days

Year	Operations	Deaths	Rate (%)
1986*	2	0	0
1987	12	4	33
1988	16	8	50
1989	12	2	17
1990	29	7	24
1991	37	3	8
1992	42	8	19
1993	45	11	24
1994	37	10	27
1995†	34	7	21
1996	29	6	21
1997	21	4	19
1998	23	1	4
1999‡	24	4	17
2000§	8	4	50
Total	371	79	21

*First operation 7 Nov 1986.

†Risk factor data available 12 Apr 1995 onwards.

‡Risk factor weightings published 15 Oct 1999.

§The last heart transplant operation at St George's was performed on 22 Oct 2000.

of the procedures are claimed to have optimal properties for the present purpose.

We applied each test retrospectively from the beginning of the heart transplant programme to determine the earliest time, if any, that the result became positive. We also applied each test at the end of the programme, by which time 371 transplants had been carried out. For example, we tested whether the mortality for all 371 cases was significantly greater than a benchmark of 15%.

False positive (type I) error

A false positive or type I error occurs when a result is positive by chance and thus raises a false alarm regarding the death rate. We evaluated the false positive rate for each test in isolation.

When a significance test using a fixed critical P value such as 0.05 is applied to a true hypothesis every time that an outcome in an unending series becomes known, then the null hypothesis—that the death rate is satisfactory—will eventually be rejected and a false alarm is bound to occur. An example of repeated significance testing in relation to child heart surgery in Bristol has been discussed.² In the absence of real changes, the type I error rate for indefinitely repeated significance tests is 1. However, a control process based on repeated significance testing can be helpful provided that the number of cases before a false positive occurs, called the run length, is large compared with the frequency with which actual changes occur.

Transplant data and the national average mortality as a benchmark

We analysed death or survival within 30 days of operation (table 1). The series of transplant cases was sequenced in the order in which the operations were carried out. We used a national average 30 day mortality of 15% as the benchmark.

Seven methods examined

Average mortality—To test whether the death rate, expressed as the number of deaths divided by the number of operations, was significantly different

from 15% we used a two tailed test at the 0.05 level of significance.

Excess mortality—The concept of excess mortality was used at the General Medical Council inquiry into child heart surgery in Bristol to argue that surgery should have stopped sooner than it did (expert opinion for the General Medical Council from D J Spiegelhalter, "Statistical analysis of surgical data provided by Bristol Royal Infirmary," Feb 1997).³ We added a margin of 5% to the benchmark of 15% to define "excess" mortality to be 20%. We tested whether the mortality was significantly greater than 20% by using a repeated one sided test at the 0.05 level of significance. The test consisted of seeing if the lower one sided 95% confidence limit for the mortality exceeds 20% at any stage.

Moving average—No calculations or special skills are required for the moving average test. We tested whether there were eight or more deaths in any 10 consecutive cases during the transplant programme.

Run of deaths—The run test is even simpler. A "run" of deaths occurs when several consecutive patients die. We tested whether there was a run of five deaths at the end of the series, as was thought to have occurred, and at any time within the series.

Sequential probability ratio test—The sequential probability ratio test has formal statistical properties.⁴ We used a benchmark failure rate of 15% with an alternative failure rate of 20% and the values of type I error as used by de Leval et al ($\alpha = 0.05$ and power $\beta = 0.20$).⁵

Cusum graph with v-mask—Samples of a process can be measured and, after the deduction of the target mean of the process from each measurement, the cumulative sum of the measurements should be approximately zero. When plotted against the sample number, the cumulative sum will therefore seem more or less horizontal if the process is in control.⁶ This can be tested by placing on to the trace a mask in the shape of a "V" that is lying on its side, so that it looks like a large "greater than" sign. The mask is determined solely by the choice of the apex angle. The horizontal distance ahead of the point, representing the latest case in the series at which the apex of the mask is to be placed, must also be specified. The test consists of seeing if all the data points lie within the arms of the mask. An equivalent mask specified by a height, h , and slope, k , requires less drawing space as it is placed on the trace at the latest point rather than some distance ahead of it (see fig 6). The test still consists of seeing if all the data points lie within the arms of the mask (see bmj.com).

Cumulative risk adjusted mortality (CRAM) chart—The cumulative difference between the expected and observed number of deaths shown on the vertical axis of the CRAM chart is the same as in the cusum plot except that the direction is reversed. Unlike any of the other methods, however, the CRAM chart allows different risks for different patients. Risk factors and association with death within 30 days are presented on bmj.com. The performance ratio at any point is estimated as the observed number of deaths up to that point divided by the corresponding expected number of deaths. Where there are sufficient data, control limits can be calculated to detect a change in the performance ratio (see bmj.com).

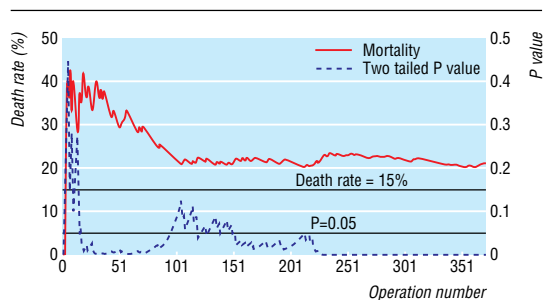


Fig 1 Average mortality in 371 heart transplantations in one hospital compared with national average (15%)

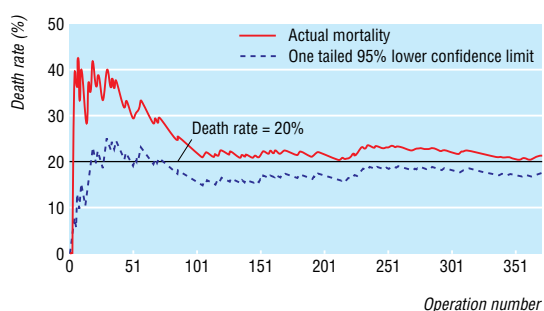


Fig 2 Excess mortality in 371 heart transplantations in one hospital compared with national average plus margin of 5%

Results

Average mortality

The death rate exceeded the benchmark of 15% from the fourth operation onwards (fig 1) but did not become significant—that is, P value below 5%—until operation number 16. For the complete series, the observed mortality was 21% ($P=0.0015$, two tailed). The probability of a type I error from repeated significance testing throughout the series is 0.17 (see table 3)—that is, this test has a false positive rate of about 1 in 6.

Excess mortality

The death rate was above 20% by the fourth operation but this was not significant (fig 2). By operation number 19 there was significant evidence of excess mortality ($P<0.05$, one tailed). At the end of the series there was no significant evidence for excess mortality.

Moving average

The death rate as a moving average of 10 operations reached 80% only once, at operation 230 (fig 3). At other times, the moving average was not significant for deaths within 30 days, as defined here, including at the end of the series when the moving average was 50%. The newspaper account of eight deaths in the last 10 cases was presumably based on a different period of survival or sequencing of cases.

Runs of deaths

The longest run of consecutive deaths was five, and this occurred only once, at operation number 230 (fig 4). There were only two deaths in the last five cases. The type I error rate for repeated examination for a run of five or more deaths in a series of 371 operations with 15% event rate is 0.023 or 1 in 43.

Sequential probability ratio test (SPRT)

At operation number 56 the sequential probability ratio test indicates that the death rate was 20% rather than 15% (fig 5). The type I error rate for the repeated test was set to 5%. Strictly speaking, the plot and the test are not relevant after one of the control lines has been crossed, because once the decision between a benchmark death rate of 15% or 20% has been taken the test does not allow for a reversal of the decision. The final point on the plot was above the 20% limit.

Cusum graph with v-mask

A truncated v-mask is shown in figure 6 at operation number 57, which was the first occasion that a mortality greater than 15% was signalled. The mask is shown again at the end of the series. If we assume no change in death rate from 15%, the average run length before a different death rate is signalled would be 3662 operations. On the other hand if the death rate increased to 20% it would be 29 operations.

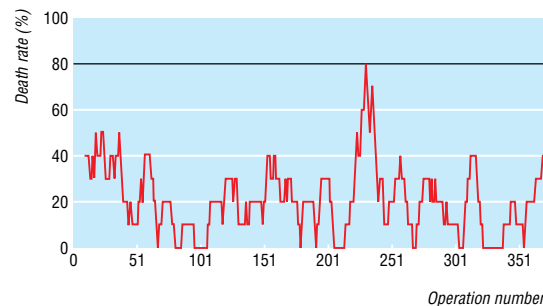


Fig 3 Moving average of 10 heart transplantations in which there were eight or more deaths

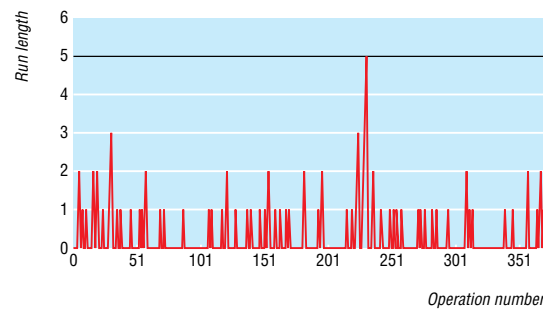


Fig 4 Run length of deaths when five consecutive patients undergoing heart transplantation died

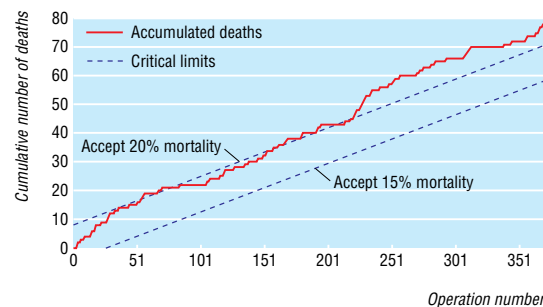


Fig 5 Sequential probability ratio test in 371 heart transplantations in one hospital

Table 2 Summary of test results to detect excess mortality in series of heart transplantations

Benchmark or test criterion	Average mortality	Excess mortality	Moving average of 10	Run of deaths	Sequential probability ratio test	Cusum with v-mask	CRAM
Benchmark or test criterion	15%	20%	80%	5 deaths	15% v 20%	15%	Internal control
Result of test at end of series	+ve	-ve	-ve	-ve	+ve	+ve	-ve
Type I error rate for test at end of series	0.05, two tailed	0.05, one tailed	0.00002	0.00008	0.018, one tailed	0.031	0.01
Result of test throughout series	+ve	+ve	+ve	+ve	+ve	+ve	+ve*
Type I error rate for test throughout series	0.17	0.28	0.002	0.023	0.05	0.11	0.25

+ve=positive result for out of control, -ve=negative result for out of control.
 *Decrease in death rate.

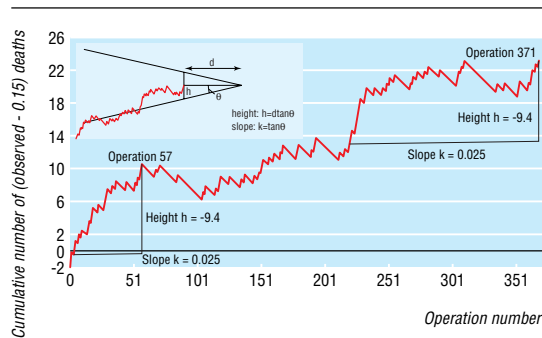


Fig 6 Cusum with truncated V-mask at end of series of heart transplantations and at first operation at which series is out of control

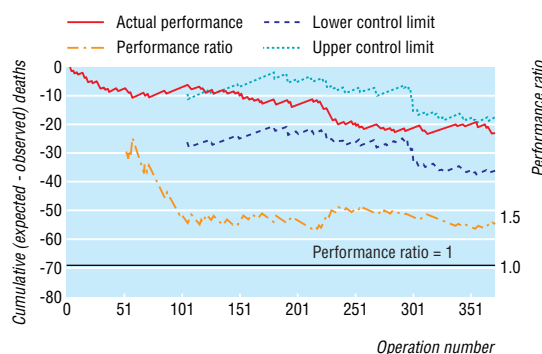


Fig 7 CRAM chart with uniform external risk estimate of 15% for all patients, showing performance ratio of observed number of deaths to number of deaths predicted by external estimate

CRAM chart (not adjusted for risk factors)

Individual target risk estimates were not available for the early part of the series, so we used a uniform external risk estimate of 15% mortality to draw figure 7. The upper control limit was crossed at the first determination of the control limits, which was at operation number 104. The test result was positive in the sense that the control limits were reached and a change in death rate was signalled, but the change was towards a lower mortality than had occurred earlier in the programme. As we did not adjust for risk factors, one reason for the improvement may have been a shift to lower risk patients. As with the sequential probability ratio test, there are some uncertainties in interpretation of control limits once they have been crossed; however, it seems reasonable to infer from figure 7 that the death rate was within the limits at the end of the series.

What is already known on this topic

Hospitals do not routinely monitor every inpatient death

It is difficult or impossible to say whether a run of deaths is significant unless prospective methods are used

National newspapers reported that the death rate for the last 10 patients who received a heart transplant at one hospital was more than five times the national average

What this study adds

Retrospective analysis of those transplant data, carried out with several different statistical process control techniques, gave different answers as to whether the death rate was acceptable at the end of the transplant programme

Deaths within hospital should routinely be monitored with CRAM charts

Summary of results

At the end of the series the average mortality, sequential probability ratio, and cusum tests indicated a level of deaths higher than the benchmark, and the remaining four of the seven statistical tests yielded negative results (table 2). Six of the tests showed that the transplant programme had a level of deaths above benchmark at some point. The point at which an alarm would first have occurred varied with the choice of method. With the CRAM chart, the only change detected was a decrease in the death rate early in the programme.

Discussion

Evaluation of the type I or false positive error rate is essential if a high death rate is to be distinguished from a run of bad luck. Some of the tests that we have described are complicated to apply. The national average may not be known (see bmj.com), and there is no guidance on what is an acceptable departure from the national average. In the latter stages of the heart transplantation programme mortality was high according to the average mortality test, sequential probability ratio test, and cusum with v-mask, but not by the excess mortality criterion or the other tests, including the CRAM chart. There are no methods by which to calculate the false positive rate, when the decision to test and the choice of test are made after poor results have been

observed. An above average death rate does not necessarily indicate a low quality of service.

Contributors: See bmj.com

Funding: NHS Executive research and development project grant number SPGS738.

Competing interests: JP is a trustee of Constructive Dialogue for Clinical Accountability (CDCA) and a member of St George's Mortality Monitoring Group and is funded by St George's Healthcare NHS Trust. PJ is medical director of St George's Healthcare NHS Trust and chairs the St George's Mortality Monitoring Group.

Ethical approval: Not required.

- 1 Booth J. Heart transplant deaths trigger official inquiry. *Sunday Telegraph* 2000 Oct 15.
- 2 Poloniecki J. Half of all doctors are below average. *BMJ* 1998;316:1734-6.
- 3 General Medical Council. Determination of the Professional Conduct Committee in the case of Mr Wisheart, Mr Dhasmana, and Dr Roylance, 18 June 1998.
- 4 Wald A. *Sequential analysis*. New York: Wiley, 1947.
- 5 De Leval MR, Francois K, Bull C, Brawn W, Spiegelhalter DJ. Analysis of a cluster of surgical failures. *J Thoracic Cardiovasc Surg* 1994;107:914-24.
- 6 Barnard GA. Control charts and stochastic processes. *J Roy Stat Soc Ser B* 1959;21:239-71.

(Accepted 13 November 2003)

doi 10.1136/bmj.37956.520567.44

Should surgeons take a break after an intraoperative death? Attitude survey and outcome evaluation

Antony R Goldstone, Christopher J Callaghan, Jon Mackay, Susan Charman, Samer A M Nashef

Abstract

Objectives To investigate attitudes of cardiac surgeons and anaesthetists towards working immediately after an intraoperative death and to establish whether an intraoperative death affects the outcome of subsequent surgery.

Design Questionnaire on attitudes to working after an intraoperative death and matched cohort study.

Setting UK adult cardiac surgery centres and regional cardiothoracic surgical centre.

Participants 371 consultant cardiac surgeons and anaesthetists in the United Kingdom were asked to complete a questionnaire, and seven surgeons from one centre who continued to operate after intraoperative death.

Main outcome measures Outcome for 233 patients operated on by a surgeon who had experienced an intraoperative death within the preceding 48 hours compared with outcome of 932 matched controls. Hospital mortality and length of stay as a surrogate for hospital morbidity.

Results The questionnaire response rate was 76%. Around a quarter of surgeons and anaesthetists thought they should stop work after an intraoperative death and most wanted guidelines on this subject. Overall, there was no increased mortality in patients operated on in the 48 hours after an intraoperative death. However, mortality was higher if the preceding intraoperative death was in an emergency or high risk case. Survivors operated on within 48 hours after an intraoperative death had longer stay in intensive care (odds ratio 1.64, 95% confidence interval 1.08 to 2.52, $P=0.02$) and longer stay in hospital (relative change 1.15, 1.03 to 1.24, $P=0.02$).

Conclusion Mortality is not increased in operations performed in the immediate aftermath of an intraoperative death, but survivors have longer stays in intensive care and on the hospital ward.

Introduction

There are no guidelines, and no real consensus about whether or not surgeons should continue to

operate in the immediate aftermath of an intraoperative death. A survey of Welsh consultant orthopaedic surgeons underlines the lack of consensus.¹ In this study only one of the 16 orthopaedic surgeons who had experienced a patient's intraoperative death decided to cancel further operations that day.¹ Given the differences between cardiac and non-cardiac surgery, Briffa has suggested that cardiac surgeons may behave differently.² Many anaesthetists feel that intraoperative death affects them equally, if not more so.³

We explored and compared the attitudes of cardiac surgeons and anaesthetists to working after an intraoperative death. We also sought to determine whether an intraoperative death has an adverse effect on subsequent operations by the same surgeon.

Methods

Questionnaire study

We compiled a database of UK adult cardiac surgery centres using the National Adult Cardiac Surgical Database.⁴ Hospitals were telephoned and asked to supply the names of all consultant cardiac surgeons ($n=198$) and anaesthetists ($n=288$). An anonymous postal questionnaire was designed to establish information about experiences of intraoperative deaths, factors influencing the decision to stop working after an intraoperative death, and opinions on proposed guidelines for working after an intraoperative death.

Outcome study

Papworth Hospital prospectively collects data on patient demographics, risk profile, operation details, and outcome in a dedicated database. All patients are stratified for risk with the Parsonnet⁵ and EuroSCORE⁶ models. There were 81 intraoperative deaths in five years during operations carried out by all seven

Editorial by de Leval and p 375

MRC Cancer Cell Unit, University of Cambridge, Cambridge CB2 2XZ

Antony R Goldstone
clinical research fellow

Department of Surgery, University of Cambridge, Cambridge CB2 2QQ

Christopher J Callaghan
honorary clinical fellow

Papworth Hospital, Cambridge CB3 8RE

Jon Mackay
consultant anaesthetist
Samer A M Nashef
consultant surgeon

MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR

Susan Charman
medical statistician

Correspondence to: Samer Nashef
sam.nashef@papworth.nhs.uk

BMJ 2004;328:379-82



This is the abridged version of an article that was posted on bmj.com on 20 January 2004 and amended on 27 January 2004: <http://bmj.com/cgi/doi/10.1136/bmj.37985.371343.EE>