

# Papers

## 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 graph 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 institutions 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."<sup>1</sup> 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: mortality above the national average, mortality excessively above the national average, a test of the moving average mortality, a test of the number of consecutive deaths, the sequential probability ratio test (SPRT),<sup>2</sup> cusum (cumulative sum) graph with v-mask,<sup>3</sup> and CRAM (cumulative risk adjusted mortality) chart.<sup>4</sup> There is some overlap of the principles behind these tests—for example, the association between SPRT and cusum with v-mask has been discussed by Basseville.<sup>5</sup> None of the procedures are claimed to have optimal properties for the present purpose.

Each test was retrospectively applied 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 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. The likelihood of a test to give a false alarm—that is, its type I error rate—is not dependent on what happens in the actual series of cases. It is a property of the test derived from theoretical series of cases in which the true death rate is a known constant. We evaluated the false positive rate for each test in isolation. We did not evaluate an overall false positive rate for all seven tests combined into a single composite test.

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 hypothesis 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.<sup>6</sup> 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.

 Detailed statistical methods of determining the false positive rate and an extra table of data can be found on [bmj.com](http://bmj.com)

**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.

### 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. The box shows the risk factors for adult heart transplantation, and table 2 shows the clinically derived data according to these risk factors from 1995 onwards. 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.

#### Risk factors in adult heart transplantation

Recipient aged > 50 years  
 Preoperative ventilatory support  
 Preoperative circulatory support  
 More than one previous sternotomy  
 Pulmonary vascular resistance > 200 dynes (2.5 Wood units)  
 Male with body surface area > 2.5 m<sup>2</sup>  
 Retransplant  
 Ischaemic time > 3.5 hours  
 Donor aged > 45 years  
 Donor inotropic support > 10 µg/kg/min dopamine  
 Female donor  
 Ratio of donor to recipient body surface area < 0.7  
 Donor with diabetes  
 History of drug misuse in donor

**Table 2** Risk factors\* in adult heart transplantation

Factor count	Risk (%)†	No of cases	No (%) of deaths
0	3	22	3 (14)
1	5	45	7 (9)
2-3	13	59	12 (20)
≥4	20	6	2 (33)
Total	9	132	24 (18)

\*See text box.

†Risk of death within 30 days.<sup>11</sup>

*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).<sup>7</sup> 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.<sup>2</sup> 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  $\beta=0.20$ ).<sup>8</sup>

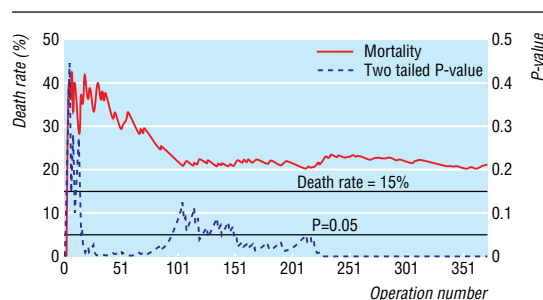
*Cusum graph with v-mask*—Samples of a process can be measured and, after the deduction of the target mean of the process from each measurement, the cumulative sum of the measurements should be approximately zero. When plotted against the sample number, the cumulative sum will therefore seem more or less horizontal if the process is in control.<sup>3</sup> 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](http://bmj.com) for further details).

*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. 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](http://bmj.com) for further details).

## 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.



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

**Excess mortality**

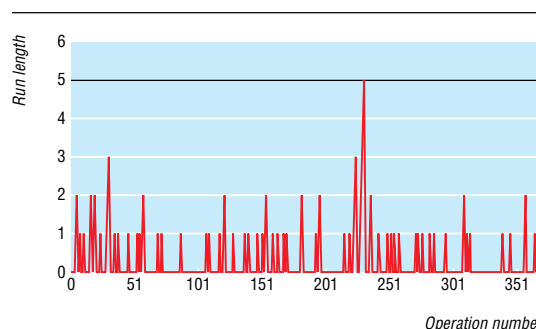
The death rate was above 20% by the fourth operation but this was not significant (fig 2). By operation number 19 excess mortality was significant ( $P < 0.05$ , one tailed). At the end of the series there was no significant evidence for excess mortality.

**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 (operation 230 being the fifth in the run) (fig 4). Only two deaths occurred 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.



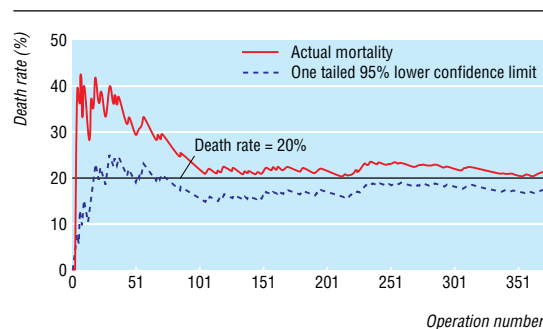
**Fig 4** Run length of deaths when five consecutive patients undergoing heart transplantation died

**Sequential probability ratio test (SPRT)**

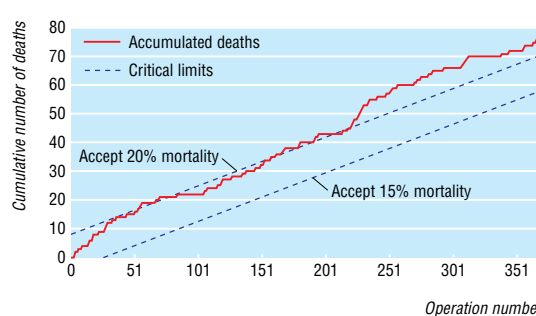
At operation number 56 the sequential probability ratio test indicates that the death rate was 20% rather than 15% (fig 5). The type I error rate for the repeated test was set to 5%. Strictly speaking, the plot and the test are not relevant after one of the control lines has been crossed, because once 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**

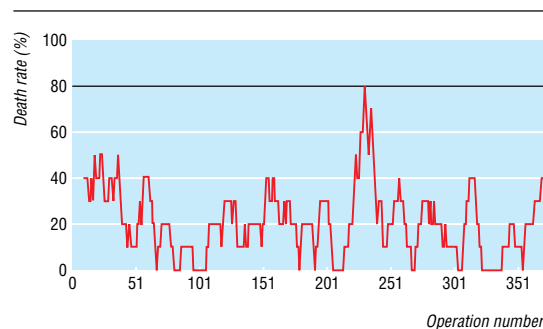
A truncated v-mask is shown in figure 6 at operation number 57, which was the first occasion that a mortality greater than 15% was signalled. The mask is shown again at the end of the series. If we assuming no change in death rate from 15%, the average run length before a different death rate is signalled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



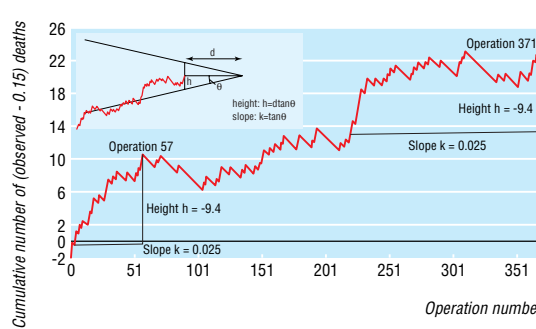
**Fig 2** Excess mortality in 371 heart transplantations in one hospital compared with national average plus margin of 5%



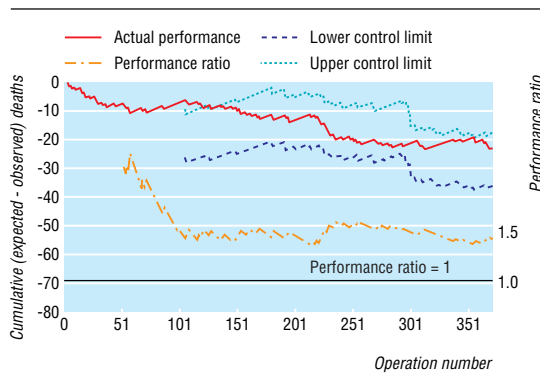
**Fig 5** Sequential probability ratio test in 371 heart transplantations in one hospital



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



**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

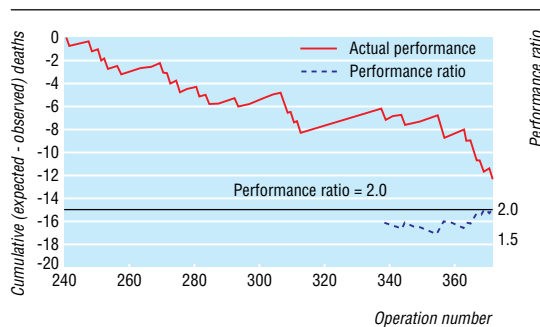


**Fig 7** CRAM chart with uniform external risk estimate of 15% for all patients, showing performance ratio of observed number of deaths to number of deaths predicted by external estimate

**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.

*Risk adjusted*—Data on risk factors (see box and table 2) were available from the 240th transplant, but we could not calculate



**Fig 8** CRAM chart with uniform external risk estimates of 3%, 5%, 13%, and 20% and showing performance ratio of observed numbers of deaths to number of deaths predicted by external estimates

control limits as there were not enough operations. The CRAM chart, however, provides prospective risk estimates for individual patients after the first 16 deaths even in the absence of control limits (fig 8). The 16th death in cases with data on risk factors was transplant number 338. For all the remaining operations the performance ratio was close to 2—that is, the observed number of deaths remained at about twice the number of deaths predicted by the risk factors in table 2.

**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 3). 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**

It is surprisingly difficult to determine a contemporary national average mortality, even for such a high profile activity as heart transplantation. In 1998 the British Transplantation Society cited 9% (95% confidence interval 5% to 13%), giving as source the UK Cardiothoracic Transplant Audit, reproduced from UK Cardiac Surgical Register 1996-7.<sup>9</sup> According to the Royal College of Surgeons of England, 12.3% of patients died within 30 days of first time adult heart transplant in the United Kingdom between 1995 and 2000 (personal communication). An estimate of 12.1% (70% confidence limits of 10% to 14%; 45/373), however, from April 1995 to December 1999, seems to be the only estimate in a peer reviewed journal.<sup>10</sup> If we deduct the operations at St George's Hospital during the same period, the average was 10.7% for the remaining centres, with 15% being the upper two tailed 99.3% confidence limit.

For most groupings of patients it is not practical for a hospital acting on its own to compare with the national average or with other units. CRAM charts, however, can be used by a hospital to monitor any grouping of its patients and to identify recent changes in mortality that merit prompt investigation, as these charts can be produced by using only local data and do not need an external benchmark.

**On-site mortality monitoring**

Without a prospective monitoring system the inevitable occurrence of poorer results, sooner or later, may lead to a damaging interruption of service and loss of confidence. It will be difficult or impossible to undo the damage by means of a retrospective external inquiry. We suggest that an internal mortality

**Table 3** Summary of test results to detect excess mortality in series of heart transplantations

	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%	5deaths	15%v20%	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.

monitoring group, comprising clinicians, senior management, and clinical audit, oversees prospective monitoring of all deaths in hospital. Although hospital episode statistics (HES) data are collected in all NHS hospitals for administrative rather than clinical reasons, the data include demographic, admission, diagnostic, and procedural information that can be used to adjust for case mix.

In addition to occasional investigations prompted by the statistical process control there is a role for routine review of deaths in hospital. The multiple logistic regression used to calculate case-mix adjustment for the CRAM charts provides an estimate of risk, on admission, for each case. Inclusion of the estimate on a monthly mortality list can highlight a death as unexpected. However, a recommendation to report “unexpected deaths during or after medical or surgical treatment” to the coroner would need careful evaluation.<sup>11</sup>

### Summary

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, 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: Andrew Murday provided all the data and encouraged us to analyse them. He commented on an early version of the paper. JP did the analyses, wrote the paper, and is guarantor. CS managed the data and provided the graphs. MB reviewed the analyses. PJ wrote the paper.

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.

### 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 is recommended

- 2 Wald A. *Sequential analysis*. New York: Wiley, 1947.
  - 3 Barnard GA. Control charts and stochastic processes. *J Roy Stat Soc Ser B* 1959;21:239-71.
  - 4 Poloniecki J, Valencia O, Littlejohns P. Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *BMJ* 1998;316:1697-700.
  - 5 Basseville M, Nikiforov IV. *Detection of abrupt changes: theory and application*. New Jersey: Prentice-Hall, 1993.
  - 6 Poloniecki J. Half of all doctors are below average. *BMJ* 1998;316:1734-6.
  - 7 General Medical Council. Determination of the Professional Conduct committee in the case of Mr Wisheart, Mr Dhasmana, and Dr Roylance, 18 June 1998.
  - 8 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.
  - 9 British Transplantation Society. Towards standards for organ and tissue transplantation in the United Kingdom. British Transplantation Society, 1998. (accessed Nov 2003).
  - 10 Anyanwu AC, Rogers CA, Murday AJ. A simple approach to risk stratification in adult heart transplantation. *Eur J Cardiothorac Surg* 1999;16:424-8.
  - 11 Death certification and investigation in England, Wales and Northern Ireland—the report of a fundamental review 2003. London: Stationery Office, 2003. (accessed Nov 2003).
- (Accepted 13 November 2003)
- doi 10.1136/bmj.37956.520567.44

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