

# FINAL REPORT

CONFIDENTIAL

## Peer Review Report

**Overview of the statistical evidence presented to the  
Bristol Royal Infirmary Inquiry concerning the  
nature and the outcomes of paediatric cardiac  
surgical services at Bristol.  
(Spiegelhalter, Evans, Aylin and Murray)**

Professor S Gallivan

October 2000

Paper No. 574



Clinical Operational Research Unit  
Department of Mathematics  
University College London

## **1. PREAMBLE**

The statistical evidence presented to the Inquiry is complex and, as far as this reviewer is concerned, an admirable job has been done by those commissioned by the Inquiry to examine relevant data and their interpretation. The overview report is clearly a key document since it brings together many disparate strands of statistical evidence, discusses the strengths and weaknesses of data sources and the analysis methods used and provides a distillation of the important messages that emerge. The authors are to be congratulated for producing such a useful overview of the complicated statistical information the Inquiry has had to consider.

The nature of the present peer review report should be clarified. As a member of the panel of experts advising the Inquiry on statistical issues, I have been aware of the many detailed analyses that have been commissioned by the Inquiry. I have also read through the various detailed reports that have been produced and have been present at several discussions of the analytical work. As such, the independence of my position may well be questionable. On the other hand, having been free from the onerous analysis tasks commissioned by the Inquiry, yet knowing the subject matter well is undoubtedly an advantage. Hopefully, this allows the overview report to be viewed from a relatively objective perspective.

The terms of reference of this review requested specific comment, based on personal knowledge and expertise, on the following issues:

- whether the analytical approach is statistically robust and fit for purpose;
- any errors or ambiguities of a statistical nature - especially in the interpretation of the statistical evidence;
- whether the overall conclusions drawn are reliable and valid.

Comment on broader aspects of the report was also invited. The structure of this report reflects these terms of reference, with three sections discussing the issues above and another discussing broader issues.

## **2. GENERAL OBSERVATIONS**

A problem that has been faced throughout the compilation of statistical evidence is that it is all intrinsically complicated. There are very few issues that are absolutely clear cut and caveats and provisos abound. The authors of the overview report have done a fine job in summarising the evidence, but little can be done to avoid the basic complexity of the matters being discussed. Even though great pains have been taken to simplify and condense the evidence presented, the overview report is still a daunting document, even for those with statistical experience. One purpose of this review is thus to draw attention to passages of the report which the present reviewer views as having key importance.

The readership of the report will not generally be statistically trained and may find some statistical notions and terminology alien. This difficulty is compounded by the

fact that the precision that many people expect from statistics is not matched by the quality of the data that are being examined. The authors are very open about the poor data quality of the main sources used, which inevitably raises the issue of credibility in relation to the statistical conclusions reached. A key question is whether one can be sure that the overall conclusions drawn are reliable and valid, given such poor data quality. This is a complex matter, which in part depends on purely technical aspects of the statistical analysis, but also involves beliefs about the quality of the data sources used, which are acknowledged as being imperfect. The discussion of this crucial issue is a major part of the present review.

### **3. IS THE ANALYTICAL APPROACH STATISTICALLY ROBUST AND FIT FOR PURPOSE?**

The main estimate used to assess performance is 'excess mortality', namely the number of deaths that occurred at Bristol over and above what would have been expected had performance at Bristol been similar to other centres in the country. This provides a sound basis for assessing performance. It is rather easier to grasp than other estimates that might have been considered, such as odds ratios or relative risks. Also, it is becoming something of a standard method for assessing surgical mortality, particularly in relation to cardiac surgery. The computational methods used for estimating excess mortality are, from a technical viewpoint, routine. Certainly, this analytical approach can be regarded as fit for purpose.

Considering the overview report as a whole, the key finding concerns the divergent performance at Bristol in relation to excess mortality between April 1991 and March 1995 for operations on the under-ones. These data are summarised in Table 6.2 of the Overview report. Without understating the importance of the other data summaries presented, the information summarised in Table 6.2 is central.

According to the analysis, overall mortality was appreciably higher at Bristol than expected. Even taking account of case mix, which is viewed as essential by many clinicians, there was appreciable excess mortality. Depending on which estimation method is used, this excess mortality could be of the order of 30 - 35.

However, a Devil's Advocate might argue that, because of data errors in the national sources, the process of accounting for case mix mortality might itself be in error. It is thus important to know whether we can have confidence in these excess mortality estimates.

In order to investigate the robustness of their findings, the authors have carried out an extensive programme of 'sensitivity analysis'. This somewhat technical term should be clarified for the benefit of non-statisticians. Given the poor quality of the data sources that the main analysis relies on, the authors have carried out many different alternative analyses. This involved examining different scenarios, making different sets of assumptions and drawing on different data sources. The technical details of this demanding analysis have quite rightly been consigned to a technical Appendix and only a summary of the results is presented in the main body of the report, in Section 6.4.3. This very important part of the analysis is summarised as follows:

The sensitivity analyses carried out included examining:

- removal of centres from all analyses where there is a high discrepancy between Hospital Episode Statistics (HES) and the Cardiac Surgical Register (CSR);
- removal of procedure Groups 2 and 3 which suffer from known coding overlap in CSR;
- increasing the mortality rate in the HES data for each centre to compensate for the 'undercount' related to open operations in the under-ones detected in the linkage study;
- an extreme, and probably artificial case, where mortality estimates were deliberately chosen to show Bristol in the best possible light.

In all cases, even with the final artificially optimistic scenario, analysis still indicated strong evidence for substantial excess mortality at Bristol.

In view of the evidence from this sensitivity analysis, one can only conclude that the analytical approach adopted was indeed statistically robust.

#### **4. ARE THERE ANY ERRORS OR AMBIGUITIES OF A STATISTICAL NATURE - ESPECIALLY IN THE INTERPRETATION OF THE STATISTICAL EVIDENCE?**

No major errors or ambiguities were detected in the report.

On a relatively minor issue, a purist might disagree with the use that has been made of statistical significance testing in this context. Statistical significance testing is more usually associated with the formal testing of hypotheses and tacitly presupposes scientifically well designed and controlled studies. While the data sources used in the analysis are undoubtedly useful, it would be stretching credibility too far to claim that routine data collection procedures within the NHS match those that would be expected in formal scientific studies. The statistical significance values cited give an indication of the degree of divergence, however they shouldn't be interpreted as representing formal scientific hypothesis testing.

## **5. ARE THE OVERALL CONCLUSIONS DRAWN RELIABLE AND VALID?**

### **5.1 Overall conclusions**

Here the term 'overall conclusions' will include both the 'Conclusions' section of the report and also points itemised in the 'Executive Summary'.

In general, overall conclusions match very well to the analyses that have been carried out. Certainly each statement made in the Executive Summary can be matched to an appropriate section of the report, or to a chart or Table, as appropriate. The majority of the Section named 'Overall Conclusions' is also well linked to preceding material in the report, however one section (Section 9.4), is rather more speculative. This section will be discussed further below.

### **5.2 Divergent outcome at Bristol**

Special attention deserves to be given to the reliability and validity of the conclusions concerning excess mortality, since this is a key issue as reflected by a passage cited in Section 9.5:

*"The single most compelling aspect of the data is the magnitude of the discrepancy between the outcomes observed at Bristol and those observed elsewhere"*

The overall reliability and validity of this is very much bound up with the issue of the quality of the data sources used to in the analysis. The authors are very open about deficiencies in the data sources used. To quote from the executive summary:

*"All data sources were flawed and no one source could be considered as representing the 'truth'."*

The heart of the matter is the extent to which these two views are in conflict - can one validly infer divergent outcome based on potentially erroneous data? Consideration of this very much overlaps with the issue of the robustness of the statistical analysis discussed above in Section 3. It also depends crucially, on what one is prepared to believe about the extent of errors in the data sources used.

Weighing up the case for the validity of the conclusions about excess mortality relies on two factors. One is purely objective and concerns the mathematical processes of the analysis. The other is more subjective and concerns the beliefs it is reasonable to hold about the extent of potential errors in the data sources used.

Dealing with the former, more objective issue. Technically, when estimating excess mortality, only three classes of information are used:

- caseload at Bristol;
- cases of death at Bristol;

- estimates of mortality rates for each of the 11 different operation types, aggregated over all comparator sites.

Thus, in examining the validity and reliability of conclusions about excess mortality, one need not be concerned about flaws and discrepancies in data sources **other than those that affect these three classes of information**. This is an important observation, since it would be wrong to infer that the analysis of excess mortality is dubious based on discrepancies in aspects of data sources that are not used in such an analysis.

Of the three classes of information, those relating to Bristol have been studied exhaustively and although there are some discrepancies, it is difficult to believe that the estimates used in the analysis are grossly in error.

Thus the validity of the main analysis hinges on the accuracy of the 11 estimated mortality rates for the 11 different operation types considered. Within the analysis of excess mortality, this is the only use made of data from the other centres. It should also be stressed that such information is not used on a centre by centre basis but is used only in aggregated form. Even if a mortality estimate from one of the comparator centres is inaccurate, this may not affect the analysis unduly, so long as the overall mortality estimate is reasonably sound

As discussed in Section 3, if one is prepared to believe that one or other of HES or CSR gives a reasonably accurate estimate for the 11 mortality rates, then the programme of sensitivity analysis, examining many alternative scenarios and assumptions, shows strong evidence of divergent performance. Indeed, to support this conclusion, neither data source needs to be consistently accurate across all operation types. So long as one or other gives a reasonable mortality estimate for each type of operation, the evidence remains strong. Further, it would not even matter if mortality estimates from both sources were wrong, so long as the 'true' mortality rate lay somewhere in between, or indeed close to either. The consistency of the evidence for divergence in all the scenarios examined in the sensitivity analysis adds confidence that the conclusions are reliable and valid.

In many respects there is reasonable concordance between data sources, and this provides some reassurance that mortality estimates are reasonably sound. Also, in terms of the Reviewers' personal experience of clinical data, while often of poor quality, it is rare to find major misrepresentation. However it is still something of an act of faith that for most of the 11 operation types, at least one of HES or CSR provides an adequate basis for estimating mortality.

This requires some subjectivity of belief, since it has not been established whether either of the two sources came close to representing the 'truth'. Indeed it is doubtful that this is a feasible thing to establish retrospectively.

That said, to deny the divergence of outcome, one would need to believe that HES and CSR **both** gave grossly inaccurate estimates for many of the 11 mortality rates, or alternatively, that case load and mortality at Bristol has been hugely misrepresented. In the opinion of this reviewer, this is highly improbable.

### 5.3 Factors that might influence outcome

Given such strong evidence that outcomes at Bristol were indeed divergent, analysis has been carried out to examine whether there were intrinsic factors, other than case mix, that could mitigate such findings. This is discussed in Sections 8.2.1 and 8.2.2, from which the following Table is cited.

<b>Factor</b>	<b>What is associated with Higher Mortality</b>	<b>How did Bristol compare to average</b>	<b>Comments</b>
Volume	Low	Lower	Explain small proportion of excess
Age at operation	Low	Higher average age	Marked divergence of practice at Bristol
Proportion of Down's	High	Lower proportion	Does not explain excess
Transfers	High	Lower proportion	Does not explain excess
Emergency Admission	High	Lower proportion	Does not explain excess
Socio-Economic Deprivation	High	No difference	Does not explain excess

It can be seen that none of these factors provides mitigation for the level of excess mortality estimated and provides further reassurance that the findings are not a statistical artefact. Again, this suggests that the analysis provides reliable and valid evidence of divergence of the outcomes at Bristol.

## 6. BROADER ISSUES

### 6.1 The Clinical Case Note Review

The Clinical Case Note Review assembled important information about peer judgement concerning the adequacy of care received. The Overview cites findings from this Review:

*“Their executive summary concludes that the care received by 70% of the children was adequate, leaving 30% whose case was less than adequate to different degrees. For just over 5% of children, it was considered that different management could reasonably be expected to have made a difference to outcome”.*

Quite properly, it is beyond the scope of the overview report to interpret the import of this in terms of whether it reflects an unsatisfactory level of care. This is a non-statistical matter that others are better placed to judge. However, as the authors point out, there is no evidence one way or the other that such a pattern is anomalous within the NHS. The Case Note Review was not carried out at other centres, thus there is no basis for comparison. Further, because there is no basis for comparison, such

information cannot be used as a means of determining whether the inadequacies in care that were identified 'explain' the divergent mortality rate.

However, the authors comment on the Case Note Review, and we are told:

*“The conduct of surgery was one of the criticised factors but was not particularly highlighted. In the stratified sample, over half the deaths (21/40) were considered to have received less than adequate care in which different management might have made, or would reasonably be expected to have made a difference in outcome.”*

While it should be noted that 21 out of 40 is a figure broadly comparable to the estimate of excess mortality, the lack of comparator information makes it unsafe to conclude that Bristol was divergent in terms of the overall adequacy of care. It is frustrating that information on such an important issue was not collected.

## **6.2 Issues associated with audit**

Section 9.4 of the Overview report (“What might have been known?”) discusses difficulties that might have faced Bristol carrying out its own audit in the late 1980s and early 1990s. The section is rather speculative and technical in nature and more “Discussion” than “Conclusion”. Points are made that there was a delay of 18 months before CSR data became available and that sample size for paediatric cases might have hampered formal statistical analysis. Both these points are valid as far as they go.

However, it could also be argued that the majority of data concerning deaths that occurred at Bristol was readily available locally, so audit could and should have been carried out using such up-to-date information to supplement the data from CSR. Also, it should be recognised that responsible audit at Bristol should have encompassed all cardiac surgery cases, not just the paediatric caseload. Outcomes for adult cardiac surgery at Bristol could and should have been an integral part of the audit process. For adult surgery, sample size is less of an issue and data available at the time indicate divergence from national mortality rates for one of the surgeons

## **6.3 Imperfections in data sources**

The main purpose of the statistical analyses carried out has been to examine patterns of clinical outcome at Bristol in comparison to elsewhere. In the course of this, evidence from both statistical and clinical experts has revealed uncomfortable facts concerning two national data sources, HES and CSR. Citing Section 9.1,

*“The two national sources, HES and CSR are admittedly imperfect. Both suffer considerably from lack of agreed operating procedures for ensuring completeness and accuracy of activity, coding and outcome results. Both the OPCS4 coding scheme and the use of non-clinical coders lead HES to be viewed with suspicion by clinicians. There are also strong concerns about variability between centres in the CSR’s coding procedures and the recording of mortality.”*

The gravity of these weaknesses is alluded to later, in Section 10.1,

*“Given the many flaws that have been identified in existing data sources, it is clear that only gross divergence could have been identified with any degree of confidence. If, for example, the mortality rate ... had been 50% higher than elsewhere rather than 100% higher, it would have been very difficult to exclude the possibility that the difference had arisen through a combination of differences in case mix, in the coding of operative procedures, and in the thoroughness of achieving follow-up data.”*

These passages are highlighted since they demonstrate that there are major problems concerned with information collection and collation within the NHS. The implications of this are potentially grave, particularly in relation to the potential routine monitoring of outcomes. The compilation of national data about NHS operation appears to be so unreliable that it is not a sound basis for such monitoring, and as the Overview points out, is only capable of detecting gross divergence. Quite rightly the Overview Report makes a recommendation in Section 10.1,

*“Existing data sources can and should be improved, for example by introducing routine linkage of HES records to national mortality records in order to confirm mortality data.”*

While agreeing that the present state of national data collection is inadequate, the present Reviewer does not agree with the view that data collection system can be rectified by relatively minor modifications. Consideration should be given to whether the whole process of national data collection should be thoroughly reviewed and perhaps reorganised.