

Institute of Primary Care
University of Sheffield
Sheffield S5 7AU
8th October 1999

Mrs Liz Baldock
Bristol Royal Infirmary Enquiry
2-10 Temple Way
BRISTOL BS2 0BY

Dear Mrs Baldock

Commentary on statistical analyses re letter from Dr Chadwick 4/10/99

Thank you for the Annexes A,B,C and D which compare neonatal deaths following open-heart surgery at Bristol Royal Infirmary with the rest of the UK and with data for selected European Centres.

Annex A

This contains analysis of paediatric cardiac mortality data from UBHT 1990-1992 and compares it to the national average year of 1991. It omits the neonatal arterial switch operation. It shows Bristol to be statistically significantly worse than the rest of the UK on a number of operations.

I have checked the chi-squared tests and they seem correct.

Comments

- 1) The data were analysed by chi-squared test without Yates' correction. Some of the expected numbers were small, rendering the chi-squared test less reliable. Some statisticians prefer the use of Yates' correction, which gives larger p-values and so the results are less likely to be significant, or Fisher's exact test for tables with small expected values. I carried out these tests, and the only change I got was for the A-V canal (age <1 year) where the mid-P two sided value from the Exact Test came out as $p=0.051$ (just above the critical level of 0.05).
- 2) It is better practice to provide estimates of the difference in proportions between Bristol and the rest of the UK and confidence intervals (say 99% because there are several comparisons). Also one should give exact p-values rather than (say) $p<0.005$.
- 3) Some allowance in the interpretation of the p-values should be given for the fact that 12 comparisons are being made. A very conservative approach would be to multiply the p-values by 12, (the Bonferroni correction) but even then most would remain significant (i.e. <0.05).
- 4) The assumption must be that the patients are comparable in prognostic factors for Bristol and the rest of the UK.

Annex B.

This compares Bristol with data from the European Congenital Heart Defects Database (ECHDD) , for years 1992-1994.

The important analyses are Figures 4-10, showing meta-analyses comparing the ECHDD with Bristol for 1993, and Figures 11-18 demonstrating the CUSUM analysis.

There are no conclusions from these analyses in the text. In particular the CUSUM charts are there largely for ease of visual inspection of the data. The analyses would appear to be sensible and valid.

Annex C

This compares Bristol 1992-1995 with the UK Cardiac Surgical Register 1992-1994. There are no data to comment upon.

Annex D contains a letter from Dr Bull to Dr Bolson (spelling?) discussing Cusums

Advice requested in your letter

i) *Are the statistical methodologies used (Chi-squared, Cusum etc) appropriate for the purpose of monitoring differences in mortality rates?*

The chi-squared test is a simple method for assessing whether differences in proportions could have arisen by chance. It is valid but has the usual drawbacks of a significance test:

- 1) the p-value is dependent on sample size,
- 2) a statistically significant difference is not the same as a clinically important one,
- 3) a lack of statistical significance does not mean that there is not a clinically important difference present.

The CUSUM chart is a well established technique for monitoring quality, especially in the manufacturing industry. It may be worth pointing out that the paper '*Quality control: an application of the cusum*' Williams SM, Parry BR and Schlup MMT . BMJ 1992, 302 1359-1361 had appeared during the period in question 1984-1995. In general it is a decision aid, and most importantly it requires a reference or target failure rate. This is easier to obtain in industry than in medicine where one can set an 'acceptable' failure rate on, say, economic grounds.

ii) *Are the analyses carried out feasible and appropriate and are the results soundly based? (some analyses are carried out on non-comparable years)*

The analysis in A compares Bristol 1990-1992 with UK in 1991 The authors argue this probably favours Bristol since the national average improved over the time period (i.e. if they had chosen 1992 for the UK Bristol would have looked worse).

The analysis in B compared Europe and Bristol only for 1993.

Thus I don't think the analysis has been compromised by the use of different time periods.

iii) *With specific reference to the period 1984-1995, were there widely available and accessible statistical methodologies that would have been more appropriate and robust for use in monitoring differences in mortality outcomes?*

I know of no other simple methodologies that might have been used and understood locally. Of course the CUSUM methodology was available before the reference period and might have been used for monitoring, but it is not a technique that was appreciated by the medical community at the time.

I hope this is of use.

Michael J. Campbell
Professor of Medical Statistics

11 October 1999

Statistical Report for the BRI Enquiry

From Klim McPherson
London School of Hygiene and Tropical Medicine

Introduction

I have Annexes A,B,C and D, which may appear under different designations elsewhere, referred to in the letter to me from Dr Chadwick dated October 4th. I am asked to comment on methodologies and the their appropriateness for measuring differences in death rates. In this context it is important to note that that question can only be answered when conditioned by the dominant hypotheses under test in these circumstances. In this case there were probably two dominant hypotheses concerning both operative processes and treatment outcomes – and these were:

1. Was Bristol different from other comparable institutions for the same type of case during the time under enquiry and
2. Were particular surgeons different from others, in general or for particular procedures.

Ancillary questions arise - to do with when any differences might have been reliably discerned and whether particular procedures might have demonstrated specific differences which were not demonstrated for others.

Annex A

This gives a rudimentary summary of some crude significance tests for the mortality differences, unadjusted for case mix, between Bristol and the rest of the UK for certain procedures. In the first set of analyses it would have been helpful to have estimated the odds ratio (OR) of death between Bristol and UK (the actual source of these UK data was not specified, was it the UK cardiac surgical register?) for each procedure. Then, if no massive heterogeneity could be demonstrated across procedures for this contrast, to have pooled the data over several procedures. That way a much more reliable estimate of a 'Bristol' effect could have been estimated. It might also have been sensible to develop a clearer prior hypothesis about the plausible effect of 'Bristol' with respect to particular procedures and to test for such an effect pooling over those procedures. Obviously the hypotheses are important and the intention of a sensible analysis should be to maximise the statistical power of the tests, while retaining theoretical consistency.

The statistical technique is widely used and is called the Mantel Haenszel test. A more quantitative technique can be used to test for differences in quantitative measurements such as bypass time and 'Extub' time using a pooled 't' test. (I could not tell whether such quantitative data were available for the rest of the UK) These methods can also be used to test for surgeon specific effects in the same manner. These analyses are, as presented in Annex A, merely explorative and should have given rise to these other investigations, given the importance of the basic hypotheses. In the pooled analyses

as presented the method of pooling is apparently crude and also incorrect, as far as I can tell.

Annex B

Tables 1 thru 3 are unexceptional, providing essentially raw data provided by Bristol to ECHDD. Table 4 could have usefully provided odds ratios of death for the two surgeons for each procedure plus a pooled estimate with a test of heterogeneity across procedures. As it stands there is too much noise from small numbers and these are under analysed, since they nonetheless suggest a clear aggregate difference. Sadly Table 5 is uninterpretable since the numbers are too few. The European data would have been useful as a generator of hypotheses. Figure 3 is likewise uninterpretable. Table 6 provides too much redundant information and should present only those statistics that are likely to enlighten – on the above hypotheses. It might then also be provided by surgeon and be pooled over procedure, unlike in Table 7. Again Tables 8 & 9 provides too much crude data and too little analysis. The hypotheses are clear, but are ignored.

Tables 10 –17 are clearly designed to impress but actually provide no useful information on the above hypotheses. The meaningful comparators are presumably between JPD and JDW. In my pack Tables 12 thru 14 are missing. But among the tables I have important differences are demonstrated on the JPD/JDW contrast. Mortality is higher for JDW while IPPV appears higher for JPD. I think control data, from other places, is essential in these analyses and I am surprised that they were not provided from the ECHDD database. Figure 4 is for all procedures and should have been more specific – and possibly pooled over index procedures as indicated above. Crucial in the interpretation is the procedure mix – about which we have no information. Analyses should have stratified for procedure. Figure 5 is devoid of information. Figures 6-9 are missing and Figure 10 is data sparse. In Figure 13 the notion of a ‘marked’ improvement in failure rate after case 54 in Europe is unsupported by appropriate analysis. In general these analyses do not conform to the usual strictures of a proper sequential analysis - by any means.

Annex C

This document introduces several ancillary hypotheses that bear on learning curves, the extant differences there may be between surgeons in outcome, leading possibly to some not performing some operations. The author also points to the lack of formal mechanisms for monitoring and attributing cause to outcomes. These analyses compare the mortality in Bristol between 1990 –1995 and those recorded by the UK cardiac register to March 1994. Presumably the Bristol data is excluded from these latter statistics.

These analyses are again very preliminary and basic. They suffer from the inadequacies stated above, but particularly they test not a single hypothesis, germane or otherwise. The data from the UK register are inadequate since they only include crude rates of mortality with no indication of the numerators. Sensible pooled analyses would be clearly indicated, as above. I simply cannot believe that the authors of this report did not have access to the actual numbers from the register, and if they

did not then the notion that there are no clear mechanisms for monitoring is simply gratuitous. The confidential and sensitive data are collected for these purposes as well as others and should obviously have been used. They were not and this omission enabled a rather complacent report based on no meaningful statistical analyses. My view is that anyone writing such a report had to rely on absolute confidentiality, since even the most basic analyses are clearly indicated given the remit of the authors – and they simply were not performed. The interpretation is thus, at minimum, subjective.

Annex D

This represents a basic manual (taken, out of context, from a GOS Cardiac Surgical Course?) on the use of Cusum methodology which might have been used (or indeed which could be used routinely everywhere, since it is cheap and easy) in Bristol. In these documents there is no evidence that it was.