Report to the Bristol Royal Infirmary Inquiry

Issues related to papers by Professor John Yates presented to the Bristol Royal Infirmary Inquiry

Professor Stephen Gallivan
8th March 2000

Clinical Operational Research Unit
Department of Mathematics
University College London
Gower Street
London WC1E 6BT

EXECUTIVE SUMMARY

1   A number of papers by Professor John Yates have been made available to the Bristol Royal Infirmary Inquiry. This report gives peer review comment that has been requested by the Inquiry. My general views have been invited, with particular emphasis on:

    i)   the appropriateness and potential value of using Health Episode Statistics (HES) data to monitor and compare trends in surgical performance;

    ii)   the statistical rigour of the analyses on which the reports are based;

    iii)   whether the conclusions drawn are scientifically robust.

    My review of the papers is summarised as follows:

2   *Examining Variation in Death Rates - a Job for the Scientist not the Journalist - by Michael Harley and John Yates, April 1994.*

Having carefully described the potential pitfalls of using NHS data for monitoring performance, the authors then brush aside these difficulties and proceed willy-nilly. A very brief results section is then used as a basis for a much longer and polemical discussion. The rhetoric of this is not matched by supporting evidence established in the paper. The paper uses rather emotive language and has an accusatory tone. Based on the evidence reported, it is by no means clear that the claims made are justified. The paper falls well short of the standards of scientific rigour that would make it suitable for publication in a reputable scientific journal. Indeed, the paper unintentionally illustrates the dangers of using routinely collected NHS data to make inferences about the performance of individual centres. The data are incomplete and inaccurate, it is difficult to standardise and, as this paper amply demonstrates, there is considerable scope for over-interpreting and perhaps mis-interpreting the information available.

3   *Early identification of poor performance and major performance failure - by John Yates, June 1995*

This is a proposal rather than a scientific report. The author was proposing that his team should examine blinded HES data for the whole of the UK to try to examine patterns of activity which appear 'inappropriate, insufficient, dangerous or inefficient'. No specific hypothesis or methodology is discussed, but rather, the proposal seems to be based on the hope that something would emerge. There are many potential pitfalls in this scattergun approach to data analysis. The proposal does not make a scientifically convincing case that the study is worth pursuing.

4   *A case study exploring the early identification of performance failure in an acute hospital - by John Yates, March 1997.*

The logic underlying this paper is flawed in that the conclusions reached are not particularly related to the hypothesis tested. Indeed the study doesn't properly test the main hypothesis, that HES data could be used retrospectively to identify the Bristol Royal Infirmary. Information from non-HES sources was needed to do this and indeed was responsible for the majority of the sifting done. It is not terribly surprising that one can identify the Bristol Royal Infirmary using such information. From this, one cannot infer very much about whether or not HES are a useful monitoring tool. The methods used are scientifically very flawed, bearing out the concerns expressed about the previous paper which seems to be the proposal that led to this study.

5.   In summary, none of the papers are convincing. There are problems with statistical and scientific rigour and the conclusions drawn are not scientifically robust.

6.   One can not infer from this that there is no value in using HES data to monitor and compare trends, however, there are sufficient doubts about its accuracy and completeness that further evidence is required to make this judgement. Although strictly speaking beyond the scope of this Review, my recommendation is that a study be carried out to assess the accuracy and completeness of the HES data. This should be based on direct comparison between a sample of HES data and a 'gold standard' source, rather than the indirect and, frankly unconvincing methods suggested by Professor Yates.

1. INTRODUCTION

1.1   A number of papers by Professor John Yates have been made available to the Bristol Royal Infirmary Inquiry. This report gives peer review comment that has been requested by the Inquiry. My general views have been invited, with particular emphasis on:

      i)   the appropriateness and potential value of using Health Episode Statistics (HES) data to monitor and compare trends in surgical performance;

      ii)   the statistical rigour of the analyses on which the reports are based;

      iii)   whether the conclusions drawn are scientifically robust.

1.2   My expertise in this area stems from considerable research experience in relation to NHS operation, numerous projects making use of data gathered in the context of health care. I have also been very active in research related to monitoring the performance of surgeons in relation to perioperative mortality, both for adult and paediatric cardiothoracic surgery. I should make it clear that I have rarely had occasion to use HES data directly and have only ever made use of data summaries from this source.

1.3   A difficulty faced in peer reviewing Professor Yates papers is that it is by no means clear what context they were written in, who the intended readership was, nor what was the original purpose in writing them. The papers predate the BRI Inquiry, so presumably they were written for some other purpose. It is not clear whether they were written as papers to be submitted to a learned journal, as articles for less formal press or as reports to clients who have commissioned research. Depending on the context, both style of expression and the level of detail about methodology would reasonably be expected to vary. Presumably, none of the papers has been peer reviewed and published, otherwise Professor Yates would have submitted the published data as evidence.

1.4   It is somewhat curious that none of the papers contain reference to peer reviewed journal articles by Professor Yates. As an academic claiming to have expertise in this area, it is surprising that such cited evidence is not present.

1.5   Given the gravity of issues being considered by the Inquiry, it seems reasonable to apply the standards of detail and evidence that would usually be expected from a peer reviewed scientific journal paper or formal grant proposal.

2. PEER REVIEW OF PAPERS

2.1   ***Examining Variation in Death Rates - a Job for the Scientist not the Journalist - by Michael Harley and John Yates, April 1994.***

2.1.1   **Summary of Review**

Having carefully described the potential pitfalls of using NHS data for monitoring performance, the authors then brush aside these difficulties and proceed willy-nilly. A very brief results section is then used as a basis for a much longer and polemical discussion. The rhetoric of this is not matched by supporting evidence established in the paper. The paper uses rather emotive language and has an accusatory tone. Based on the evidence reported, it is by no means clear that the claims made are justified. The paper falls well short of the standards of scientific rigour that would make it suitable for publication in a reputable scientific journal. Indeed, the paper unintentionally illustrates the dangers of using routinely collected NHS data to make inferences about the performance of individual centres. The data are incomplete and inaccurate, it is difficult to standardise and, as this paper amply demonstrates, there is considerable scope for over-interpreting and perhaps mis-interpreting the information available.

The Introduction

2.1.2   Turning to more detailed comments: On page 1, line 2 we are told

> *Clinicians, researchers and managers are all hesitant to see the publication of any sort of 'league table' because of the dangers of failing to compare like with like.*

This is a very sweeping statement. I don't know of any evidence to support this rather contentious view and it seems to be just a personally held belief on the part of the authors. Potential failure to compare 'like with like' certainly isn't highest on my list of worries about league tables, I have much greater concerns.

2.1.3   There are good observations made in sections entitled: Inadequacies in the data used, Differences between the patients selected and Variations in the level and adequacy of resources. This is not an exhaustive list of potential pitfalls and many other problems could in principle exist. For example, there may be selective reporting of data which may lead to bias. Decisions not to do surgery may have as much effect on excess mortality as surgery that has failed, yet the magnitude of this effect is difficult to fathom from NHS data sources. However, the authors should not be criticised for not providing an exhaustive list of pitfalls, since that was not the prime purpose of their paper.

2.1.4   We are told in the final paragraph of page 2 that

> *... it is rare for hospital professionals to openly discuss the adequacy or competence of staff.*

I'm not sure how much the authors inter-relates with clinicians, but my own experience is very much the reverse. Clinicians are renowned for talking shop incessantly and discussions about other colleagues are commonplace. True, this isn't always done against a background of formal statistics.

2.1.5   Again in the final paragraph of page 2, we are told

> *Any attempt to discuss such variations is usually drowned under protestations about the inadequacy of the data used and the failure to standardise data in a meaningful manner.*

The validity of this is questionable and the emotive and accusatory tone inappropriate. This is a particularly inaccurate portrayal of cardiothoracic surgeons. Certainly, it has been acknowledged that audit is difficult and that standardising data and adjusting for case-mix are big issues. However, the profession appear to have been very solidly behind audit and quality assurance as evidenced by the extensive audit work in the field and work by many researchers.

2.1.6   The rationale in first paragraph of page 3 is very difficult to follow. Having pointed out all the potential pitfalls of NHS data, we are told

> *Our experience of NHS data in other fields, eg the study of waiting lists and examination of surgical workload, suggests that despite the inaccuracies of the databases to hand, they are usually sufficiently robust enough to identify significant differences in performance.*

This is a central issue and yet this justification is very flimsy with no cited evidence to support this belief. The reader is being asked to take on trust that the NHS data they have used are sufficiently accurate and unbiased to be used as a basis of a scientific study of mortality rates. But it appears to be wholly a matter of trust, since the support they give for this view, unreferenced studies of workload and waiting lists, doesn't seem to have relevance to the issue of mortality. This is a major scientific criticism, because if the databases are indeed sufficiently inaccurate, incomplete or biased, and the authors have acknowledged that some think this may be the case, then the conclusions from the study would be worthless. It is surely not reasonable to expect the acceptable quality of the NHS data to be taken on trust. The experience of this reviewer as regards data collected about NHS operation, is that in general they are indeed of dubious quality.

The Methods Section

2.1.7   In the first paragraph of the Methods section, we are told that 6% of records did not have diagnostic or operative information. It would have been useful to know about this missing information and whether the proportions missing varied systematically

between centres. It would also been useful to know what proportion of this 6% was recorded as deaths. If it is a high proportion, then there would be cause for concern about the analysis, given the overall death rate of 2.6% later reported. Missing over twice the information as there are deaths is in itself rather worrying.

2.1.8 It is not clear in the Methods section how account was taken of the emergency status of patients. Emergency operations would certainly be expected to have a higher than average operative mortality. There are many reasons for this that would vary from centre to centre. Emergency status at admission is different from emergency status when operated upon. For example, a elective admission may develop a bleed that needs emergency surgery.

The Results Section

2.1.9 In the results section, we hear that the majority of deaths (60%) died on surgical wards without having had a surgical operation. This is certainly an important statistic to bear in mind if one is considering the possibility of using such NHS data to detect inadequate surgical performance.

2.1.10 In paragraph 2 of the Results section. The use of the terminology 'the average team' is very misleading. This is rather like saying that the average gender is half male and half female.

2.1.11 Tables 1 and 2 devote special attention to 'Team A' and 'Team B', yet nowhere in the Methods or Results sections are we told how these teams were chosen, nor the rationale for so doing. Presumably they are the 'best' and the 'worst'. The laws of mathematics dictate that if you have a list of 128 numbers, there will be a smallest and a biggest. Are we to infer much more than that about these teams?

2.1.12 The final paragraph of the Results section highlights the relatively high mortality of vascular surgery (presumably the authors mean relative to other forms of surgery). This is not terribly surprising since breakdown of the blood supply system is a very common cause of death.

2.1.13 In the final paragraph of the Results section, the authors state that 'there is still a relatively wide variation within the vascular team cluster'. There seems to be no evidence presented that supports this claim. The variability shown in Figure 2 indicates that the variability between the vascular group is comparable to that for the non-vascular group. There is a danger here that the authors are using their own value judgements about what is reasonable or unreasonable in terms of variability.

2.1.14 Several times within the Results section, the authors use the word 'crude'. They refer to 'crude hospital death rates', 'crude information not standardised for case mix information or even age of patient' and 'relatively crude hospital episode data'. Yet the fact of the matter is their results don't actually extend beyond this first crude stage of analysis. At the very least one would expect some attempt to standardise death rates for age, emergency status and oncological status. Absence of this makes the paper little more than a preliminary descriptive analysis, certainly not a sound basis for scientific inference.

The Discussion Section

2.1.15 The whole tone of the discussion is one of invective written in a journalistic style. This is ironic, given the title of the paper. With the relatively sparse amount of factual information reported within the paper, and the many shortcomings from a methodological point of view, there is little within the paper to support the views expressed.

2.1.16 In the first paragraph, the NHS is asked to explain why one surgical team had a death rate of 6.7%. The figure cited in the text is 6.1%.

2.1.17 Within the paper, little has been reported about the team singled out for attack. Of the order of 60% of their deaths possibly died before being operated on (See 2.1.9) which is more likely to be because their patients are sicker than most rather than surgical incompetence. It is also clear that the team does a lot of vascular surgery, which itself has a high death rate. From the 10 centres who did so much surgery, the team were singled out as the one having the highest death rate, yet one of the 10 had to be in this position, just as 50% of the population have to have below average IQ.

2.1.18 It is not possible to judge from the data presented in the paper whether the team in question were actually delivering a substandard level of performance. Indeed, sadly, the paper unintentionally illustrates the dangers of using routinely collected NHS data to make inferences about the performance of individual centres. The data are incomplete and inaccurate, it is difficult to standardise and, as this paper amply demonstrates, there is considerable scope for over-interpreting and perhaps mis-interpreting the information available.

2.2 *Early identification of poor performance and major performance failure - by John Yates, June 1995*

Summary of Review

This a proposal rather than a scientific report. The author was proposing that his team should examine blinded HES data for the whole of the UK to try to examine patterns of activity which appear 'inappropriate, insufficient, dangerous or inefficient'. No specific hypothesis or methodology is discussed, but rather, the proposal seems to be based on the hope that something would emerge. There are many potential pitfalls in this scattergun approach to data analysis. The proposal does not make a scientifically convincing case that the study is worth pursuing.

2.2.1 The proposal isn't written in the style of a paper for publication, thus should not be judged in such terms. The paper does not set out new data or information nor, to be fair, is that the purpose of the paper. Given this, it is inappropriate to provide a detailed. section by section critique.

2.2.2 The style of the paper is discursive, with many references to other authors' publications. Curiously, there is no reference to peer reviewed papers by the proposer

in this field, other than his PhD thesis. An academic who claims authority in a particular field would usually cite several examples of his published work, if only to establish his credibility. This is particularly true for proposals for funding.

2.2.3   There is a slight technical inaccuracy in the reference to 'the more statistical catastrophe theory of Zeeman'. Catastrophe theory was developed by the pure mathematician René Thom, as I'm sure Chris Zeeman would cheerfully acknowledge. It concerns a field of mathematics called differential geometry. Many mathematicians would be surprised to hear it described as a being a branch of statistics.

2.2.4   The final paragraphs of page 3 attempt to make the case for the author's approach to data analysis, but no references are cited to works in peer reviewed articles, so it is difficult to judge whether the claims made are reasonable.

2.2.5   The paper ends with a proposal that his team should examine blinded HES data for the whole of the UK to try to examine patterns of activity which appear 'inappropriate, insufficient, dangerous or inefficient'. No specific hypothesis or methodology has been detailed. The proposal seems to be based on the hope that something would emerge. There are many potential pitfalls in this scattergun approach to data analysis. The approach seems dangerously close to what statisticians refer to as 'data-dredging' or 'indiscriminate analysis'.

2.2.6   With lack of detail about what hypothesis is to be tested or what methodology will be used, the proposal does not make a scientifically convincing case that the study is worth pursuing.

2.3   *A case study exploring the early identification of performance failure in an acute hospital - by John Yates, march 1997.*

2.3.1   **Summary of Review**

The logic underlying this paper is flawed in that the conclusions reached are not particularly related to the hypothesis tested. Indeed the study does not properly test the main hypothesis, that HES data could be used retrospectively to identify the Bristol Royal Infirmary. Information from non-HES sources was needed to do this and indeed was responsible for the majority of the sifting done. It is not terribly surprising that one can identify the Bristol Royal Infirmary using such information. From this, one cannot infer very must about whether or not HES data are a useful monitoring tool. The methods used are scientifically very flawed.

The title

2.3.2   This is a very misleading title. This isn't a study exploring early identification of performance failure, it is a study concerning retrospective identification of a centre already identified as suspect from other evidence sources.

The Summary Section

2.3.3   The basic premise of this study is flawed. The only hypothesis tested is that one can use HES data, amongst other things, to identify the Bristol Royal Infirmary. It is not surprising that this can be done. However, given the truth of this hypothesis, the conclusions reached about the future use of HES data do not follow as a scientific consequence, they are merely matters of opinion. If this logic were followed, one might equally argue that the columns of Private Eye were a very good predictor that something was amiss in Bristol and thus infer that the NHS should grant the magazine major funding to carry out its future quality assurance.

The Introduction

2.3.4   It should be noted that much of the background material used in this paper appears in the previous paper, '*Early identification of poor performance and major performance failure - by John Yates, June 1995'* - indeed whole paragraphs are identical. That proposal and this subsequent report are clearly linked.

2.3.5   The second paragraph refers to a weakness of retrospective analysis of events surrounding a major failure. Yet is this not the case with the present study. After the event, it has been extensively reported that the Bristol Royal Infirmary treated a lot of children and that the mortality rate was high. It is thus not surprising that one can retrospectively identify the Infirmary based on this knowledge. It is less clear, and certainly not established in this paper, that one could use HES data as a prospective scanning mechanism to warn of performance failure.

Methods

2.3.6   The first sentence is misleading. Correctly identifying the Bristol Royal Infirmary does not test the hypothesis that HES data could identify a performance failure.

2.3.7   The final sentence of the first paragraph tells us that in the enquiries that had occurred up to 1995

> *There was no published evidence that Hospital Episode Statistics data was referred to at any time.*

This may well be the case, however, it is undoubtedly true that the data sources used (internal data and the UK Cardiac Surgery Register) overlap to a large extent with HES data. The record of a death that has been passed to the HES and also to the Cardiac Surgery Register is the same item of data. The fact that it has been accessed by Professor Yates via HES does not make it new or different data. It is thus not surprising that death counts from the HES and Cardiac Surgery Register should be linked. They are highly statistically dependent on one another.

2.3.8   It is important to note that the research team have used other sources of information *'to ascertain what characteristics, if any, might distinguish the Trust from others in England'*. This goes beyond the terms of the hypothesis stated in the first line.

2.3.9 Paragraph 4 states that HES are not able to link data for all units. Does this not conflict with the view that HES data can be used to identify performance failures. What if these failures occur in the units for which data can't be linked?

Results

2.3.10 The first paragraph shows the extent to which information beyond the scope of the HES has assisted the research team in identifying the Bristol Royal Infirmary. We are told that the team had found that the unit:

> a)   *specialised in paediatric cardio-thoracic surgery;*
> b)   *has a higher than average death rate;*
> c)   *had a number of fatalities following certain 'switch' operative procedures;*
> d)   *in October 1993 stopped doing neonatal switch operations for a period*

Given this, it is almost surprising that the team felt the need to use HES data given that they already had so many clues.

2.3.11 The first sentence of the second paragraph states that the team 'refined' its hypothesis after the start of the study. This may be well intentioned, but it is extremely bad science. Hypotheses should remain fixed during the course of a study.

2.3.12 In the second paragraph, we are told that of the (over) 60 centres who admit cardio-thoracic patients each year, only 11 or 12 admit more than 100 children in any one year. Thus although the study purports to be testing whether HES data can identify the identity of the Bristol Royal Infirmary, means external to HES have been used to carry out the majority of the sifting process by excluding some 80% of the centres. This is hardly a reasonable test of the powers of divination of the HES data.

2.3.13 Paragraph 3 indicates more uncertainty with the HES data and reveals that the team were not even sure they were looking at data from the same units or not.

2.3.14 Paragraph 4 discusses more detective work making use of information beyond HES to narrow down the search. This again invalidates the hypothesis being tested.(whether HES data could be used to identify the unit).

Discussion

2.3.15 In the final sentence of the first paragraph

> *Therefore if this study has failed to identify the Bristol Royal Infirmary correctly there should be some doubt about HES data recording, both in that Trust and quite possibly in other Trusts*

Be that as it may, if the reverse were the case, then this in itself would not be evidence one way or the other that HES data recording is adequate.

2.3.16 The issues raised in the second paragraph are indeed issues, but they don't arise because of the correct identification of Bristol Royal Infirmary in this study. This study examined the hypothesis of whether HES data could be used to identify Bristol Royal Infirmary. In the event, information from other sources had to be used and indeed was responsible for the majority of the detective work that was carried out. Far from being convincing evidence that HES data should play a central role in future monitoring, this almost argues the reverse.

3. DISCUSSION

3.1   In summary, in the opinion of this reviewer, none of the papers are convincing. There are problems with statistical and scientific rigour and the conclusions drawn are not scientifically robust.

3.2   One cannot infer from this that there is no value in using HES data to monitor and compare trends, however the case Professor Yates has made is not convincing.

3.3   Anecdotal evidence is that the HES data is incomplete and has a considerable number of inaccuracies, particularly in relation to diagnostic coding.

3.4   A recent study based on HES data related to the rise in emergency admissions acknowledged misclassification problems. Reanalysis of 1995/96 data showed 2.1% more emergency admissions than previously reported, presumably because errors had been detected and removed from the data. This may sound a small change, however it is comparable to the overall annual rate of increase in emergency admissions (2.6%) that the study had been set up to investigate. This illustrates that there can be difficulties placing too much reliance on HES data and that the scale of these difficulties depend very much on the purpose to which they are to be put.

3.5   In statistical terms, sample size considerations make comparisons of performance between units a difficult task. It is not yet clear whether it is sensible to use HES data to monitor and compare trends in performance and this relies very much on how complete and how accurate the HES data are. Although strictly speaking beyond the scope of this review, my recommendation is that a study be carried out to assess the accuracy and completeness of the HES data. This should be based on direct comparison between a sample of HES data and a 'gold standard' source, rather than the indirect and, frankly unconvincing methods suggested by Professor Yates.