

Report to the Bristol Royal Infirmary Inquiry.

Key directions for the future in monitoring clinical performance

A discussion paper

Professor Steve Gallivan

November 2000

Paper No. 579



Clinical Operational Research Unit
Department of Mathematics
University College London

The research on which this report is based was funded by the Bristol Royal Infirmary Inquiry. All views expressed in the report are the responsibility of the author alone and do not necessarily represent the views of the Bristol Royal Infirmary Inquiry.

© Crown Copyright 2000

Executive Summary

1. The report discusses general issues relating to clinical monitoring in clinical contexts beyond the confines of paediatric cardiac surgery.

2. There is considerable heterogeneity in the many health care processes that constitute NHS operation and no rationale has been established concerning which might be considered as candidates for monitoring. It is also far from clear that it is sensible to seek general principles related to clinical monitoring that are applicable to all clinical activities.
3. There are also circumstances where one might doubt the apparently self evident truth that clinical monitoring is a 'good thing', or that monitoring outcomes is always either a sensible or practical option.
4. Drawing an analogy with quality control in manufacturing industry, it is argued that effective clinical monitoring should not be seen as a useful end in its own right. Its relationship with other factors affecting health care processes should be considered. These include the time it takes, the availability of staff to carry it out and health economic factors.
5. Recommendations concerning clinical monitoring should be based on an appreciation that in many aspects of health care, basic information is simply not available to carry out such monitoring. Examples are concerning outpatient care of the chronically ill and general practice, where clinicians may not have accurate information available even about basic outcomes such as death.
6. Given the heterogeneity of health care processes, there are numerous outcome or performance measures that might be monitored. To identify what these are, and to determine which should be monitored is a major exercise in its own right.
7. A major issue concerns whether monitoring should be based on observational data, collected during routine clinical practice, or non-observational data, collected as part of specially designed prospective studies. While the purist might favour the latter, resource considerations and pragmatics probably dictate that the clinical monitoring should be based on observational data.

8. Although some statistical methods are available for the analysis of observational data, this is a relatively neglected part of statistics. If widespread monitoring of clinical performance is to be introduced, and if this is largely based on the analysis of observational data, then there will be a need for considerable amount of further research to develop appropriate analysis methods for such data.
9. Statistical methods exist to assist the monitoring of adverse events. If broader based systems are to be introduced, then research into case-mix factors and their relationship to adverse event rates is a high research priority.
10. A major issue in relation to performance monitoring concerns deciding what constitutes acceptable performance. It is by no means clear who should be the arbiters of what is acceptable and certainly this is not an issue that can be resolved purely on statistical grounds.
11. Problems encountered in the use of Hospital Episode Statistics (HES) to assess performance in cardiac surgery, where the clinical outcome of death seems clear cut, suggest that it would be a wholly inadequate basis for clinical monitoring in broader areas of the health sector, particularly when outcome measures are not death. Consideration should be given to whether the whole process of national data collection should be thoroughly reviewed and perhaps reorganised.
12. While it seems natural to think of performance monitoring in terms of analysis of outcome, there are circumstances where it might be more appropriate to examine variations in clinical practice.
13. The perceived need for clinical monitoring needs to be balanced against any deleterious knock-on effects it might have on health care processes, in terms of reducing the staff time and other resources available for health care provision. In particular, the health economic impact of introducing monitoring schemes needs to be considered.

Contents

1. Introduction
2. Background
3. Preamble
 - 3.1 Expectations in relation to the quality of health care
 - 3.2 An analogy between health care processes and manufacturing industry
 - 3.3 Emergency resuscitation - the pragmatics of quality assurance
4. The paucity of data within the NHS

5. What outcome measures should be used?
6. Statistical methods to assist the monitoring of outcome
 - 6.1 Observational versus non-observational data
 - 6.2 Issues associated with the analysis of observational data
 - 6.3 Analysis of observational data concerning adverse outcomes - Cusum methods
 - 6.4 The arbitrary nature of statistical significance testing
7. What constitutes acceptable performance?
8. Inherent difficulties in pure reliance on statistical methods
9. How much data should be collected and by whom?
 - 9.1 Existing national data sources
 - 9.2 More general data for monitoring clinical performance
10. Is monitoring of clinical performance always desirable?
 - 10.1 Interaction between monitoring and other health care processes.
 - 10.2 Should clinical monitoring be based on outcomes or process?
 - 10.3 Health economic consequences of performance monitoring
11. References

1. INTRODUCTION

The commissioning brief for this report asked for advice on broad priorities for improving the analysis and use of data for the purpose of monitoring clinical performance. This advice was requested to help to inform the Inquiry Panel's understanding for possible directions for the future with the rider that the advice should distil *'key ways to strengthen statistical aspects of the routine monitoring of clinical performance in the health sector'*.

To assist the process, the commissioning brief enclosed a report of expert clinical advice on future directions for monitoring performance in the field of paediatric cardiac surgery, submitted to the Inquiry by Mr Jaroslav Stark¹. While specific comment on that report was not requested, it was suggested as a useful springboard for wider comment on clinical outcomes monitoring in the health sector as a whole. Indeed it has proved to be, since it provides a fine summary of potential recommendations for monitoring in the speciality of paediatric cardiac surgery. The present author has little to dispute as regards Stark's report, although potential bias should be declared since Stark is a research collaborator.

The challenge in the present report, is to go beyond the questions 'what can and should be done about monitoring in paediatric cardiac surgery' to consider options concerning monitoring in the health sector as a whole. This is challenging indeed. The subject matter and the questions being asked are hugely complex. It has been difficult enough to consider options for the single specialist field of paediatric cardiac surgery, a rather specialised part of the health service. To move from this to consideration of monitoring across the broad range of the health sector is more difficult by several orders of magnitude.

Within the scope of a brief report, it is not possible, nor indeed appropriate, to provide detailed technical recommendations about what should or shouldn't be done in terms of large scale and broad reaching clinical monitoring. Indeed, it is not clear that anyone is in a position to make such recommendations, nor that evidence has yet been assembled of which to base such views. The focus of the Inquiry has been on evidence related to paediatric cardiac surgery. Issues and experience related to monitoring in that field are different from those related to monitoring outcome in adult cardiac surgery and both are very different from

say, monitoring the long term care of the chronically ill, or monitoring the effectiveness of general practitioners.

There is considerable heterogeneity in the many health care processes that constitute NHS operation and no rationale has been established concerning which might be considered as candidates for monitoring. It is also far from clear that it is sensible to seek a 'common denominator' of general principles related to clinical monitoring applicable to all clinical activities. Indeed, as will be discussed, there are circumstances where one might even doubt the apparently self evident truth that clinical monitoring is a 'good thing', or that monitoring outcomes is always either a sensible or practical option.

Faced with such complexity and the lack of a basic platform of systematic evidence, this report is necessarily somewhat anecdotal and is very much a collection of observations on the general topic of clinical monitoring. Specific recommendations are few and are outnumbered by words of caution. The report is more concerned with raising issues that seem, to the author, to be important for the Inquiry panel to take into account in their deliberations.

2. BACKGROUND

The commissioning brief for this report asked that it should be based on the author's personal knowledge and experience and it is perhaps worth summarising this to put the report into context. The author's field of expertise is Operational Research. This is a broad based subject that involves the application of mathematical modelling, statistics and computer techniques to practical problems associated with effective organisation and operation of complex systems. Operational Research originally developed in the 1940s to assist with organisational and logistical tasks associated with warfare. It has since found application in many other area of human endeavour, particularly in industry. Although originally trained as a pure mathematician, the author has been involved in Operational Research for some 30 years, in a variety of contexts including industry, transport and, more recently health care.

The author is currently Director of the Clinical Operational Research Unit. This has the remit of applying Operational Research techniques in collaboration with clinicians to help to develop and evaluate new treatments and to improve the clinical management of patients. In

the course of work on health care, the author has been involved in projects in a wide variety of clinical areas including: primary health care, screening, anaesthetics, cardiology, cancer, rheumatology, pathology, fetal medicine, intensive care and gynaecology. Much of this work has been involved with evaluation and clinical outcomes, the latter interpreted in a rather general sense. Relevant to the Inquiry, research topics have included the monitoring of patient status and clinical outcome^{2,3} and investigation of methods for detecting patterns of poor performance^{4,5,6}.

It is worth stressing that while familiarity with statistical methods is essential in Operational Research, the subject is very different in nature from statistics, where the main focus is on data collection and analysis methods. Operational Research is more concerned with developing an understanding of the interaction of key elements that govern a system's operation. While data analysis may play a part in this, other non-statistical methods are frequently used. As a result, the author's observations may well come from a different perspective from those of specialist statistical colleagues. Equally, in relation to issues that are specific to statistics, opinions of such colleagues may well have more authority.

3. PREAMBLE

When considering general issues associated with the monitoring clinicians, particularly when the focus broadens to consider clinical fields other than paediatric cardiac surgery, it is important to bear in mind some of the realities of NHS operation.

3.1 Expectations in relation to the quality of health care

In the past, many outside the field of health care seem to have had a rather rose-tinted view of the NHS and how it operates. The vision of clinicians as saintly figures blessed with perfect skill in a vast range of clinical techniques, may well be suitable for soap operas, but does not reflect fact. It is undoubtedly true that health care professions attract some of the most dedicated, hard working and able individuals and they deserve the greatest of respect and admiration. However, it would be mistaken to infer from this that health care processes operate in an ideally efficient and effective fashion. NHS structure and operation is

enormously complicated and, as in all complex organisations, mistakes occur. Some individuals, or centres, may be responsible for more mistakes than others, and measures of performance may well vary. In a very real sense, this is inevitable. However much policy makers, the press or the general public wish it were otherwise, there will always be a certain number of clinical errors and there will always be variation in performance. The question is not how to eradicate such deficiencies, but how best to manage the problem.

3.2 An analogy between health care processes and manufacturing industry

An Operational Research view of health care systems, which some may find unpalatable, is to draw an analogy between clinical processes and industrial processing. In reality, hospitals are staffed by a dedicated work force who labour long hours tending for the sick, infirm and dying. In an attempt to understand the key processes and their interaction, the Operational Researcher will often set aside the human and emotional issues associated with pain, disease and death, and effectively consider a hospital as a 'factory' that processes 'widgets' (patients). The processes that go on in this 'factory' are indeed complex and there are many interactions and inter-relationships, possibly many more than is common in industry. Even so, the analogy is there. Widgets of many sorts are delivered to the factory, they undergo a range of tests and processes and while there may be some wastage, most are eventually dispatched. This very dispassionate view of the activities of a hospital, or indeed of the whole NHS, as analogous to an industrial processing plant allows one to consider health care processes from a different perspective. It provides a useful metaphor that will be used several times in the report.

One example where it can be evoked concerns the issue of quality control. In real life industrial processing, quality control is a major issue. The Director of a factory making washers, for example, will be well aware that, although the aim might be to produce perfectly formed washers each of an identical size, variations in size will occur. This might be due to a variety of reasons such as differences in raw materials or processing. If the variations in washer size are too extreme, the product may not be saleable, thus quality control, both during the manufacturing process and at its final stage might be considered to be a useful part of the factory's activity. However, a variety of questions naturally arise. How much investment should there be in quality control? What degree of standardisation should be aimed at? What methods are available for monitoring quality? What are their relative merits

and what are their relative costs? How accurate or thorough should the quality control processes need be? Such questions are very much in the province of industrial Operational Research.

The analogy with health care systems is clear. For example, when considering quality of care, it is as unrealistic to expect a hospital to eradicate all instances of substandard care as it is for the factory manager to spend his entire operating budget to ensure at all washers leaving the plant are precisely the same size. Investment in 'quality' must be weighed against other demands. In the case of the factory, washers must be produced and sold at a rate that makes the business viable. Elaborate quality control procedures may well improve the standardisation of the product, but if this is costly, or slows production unduly, then more rudimentary quality control measures might, quite rightly, be adopted. Issues such as this are very common in industry, illustrating the necessity of taking account of many inter-related factors when making decisions about one element of a manufacturing process.

This inter-relationship between difference processes and their combined overall effect on system operation has a clear analogy in health care. While it seems self-evidently true that everything should be done to ensure the highest quality standards in health care, there are circumstances where this may be infeasible, impractical or too costly. Introducing monitoring mechanisms and other methods for ensuring adequate standards, incurs costs. Not only are there monetary costs, for computers, equipment and ancillary audit staff, but there are also costs in terms of the time required by health care professionals. Every hour a clinician spends entering or checking data as part of a quality monitoring scheme, is an hour that cannot be spent in the process of health care. This is not to say that monitoring shouldn't be done, but its introduction and design cannot and should not be considered in isolation from other processes that are necessary in the health care processes. Like it or not, health care provision must be carried out within a fixed budget and make use of the work force available.

3.3 Emergency resuscitation - the pragmatics of quality assurance

It is useful to cite a concrete example from the author's own research experience.

The survival chances of an inpatient who undergoes a cardiac arrest are very much dependent on the time taken for the emergency resuscitation team (the 'crash team') to arrive once they

have been alerted. As a result, systematic monitoring of crash team arrival times might well appear to be a high priority, yet it is difficult to do with any great accuracy. Prior to the crash team's arrival, ward staff will have been doing what they can to help the patient and cannot be expected to defer this in order to note the exact times that events occur. Once the crash team arrives, they are far too busy to take time to compile accurate records. At best, audit records will be recorded after the event and will be very prone to error, particularly as regards time estimation, given the stressful circumstances of emergency resuscitation. In such cases, accurate monitoring of response time is not practical unless at inordinate cost.

Even without accurate monitoring of crash team arrival times, it is known that success rates for emergency resuscitation could almost certainly be improved by making sure that the entire nursing staff of a hospital are trained in additional emergency resuscitation skills beyond the basic life support methods which are part of their routine training. What is a Chief Executive to do faced with the reality that the majority of her Trust's nursing staff do not have such resuscitation skills? Introducing a policy of only recruiting nursing staff with such skills would be folly. The national shortage of nursing staff is already close to crisis levels, and recruitment is difficult enough even without imposing additional conditions. A Trust-wide training programme, even if feasible, would be inordinately expensive and would itself pull nurses away from direct patient care during their additional training. Chief Executives have little option other than to accept the situation as it is, as regards nurse training levels, in spite of the fact that better quality provision could be available at a price.

As this illustrates, it is necessary for hospital managers, and indeed policy makers, to adopt the pragmatics of industry whether implicitly or explicitly.

4. THE PAUCITY OF DATA WITHIN THE NHS

It is ironic that the Bristol Inquiry should be focussed on activity in paediatric cardiac surgery. Not only is this a relatively small and very specialised field of health care, but cardiac surgery is at the better end of the health care spectrum as regards audit and methodical record keeping. Also, the short term nature of the principle outcome assessment, death within 30 days, makes it relatively easy to assess. Even so, the statistical studies

commissioned by the Inquiry give evidence of substantial shortcomings in the quality and completeness of data gathered about caseload and outcome⁷.

When considering questions associated with monitoring in clinical fields beyond cardiac surgery one should take account of the fact that, as regards methodical data recording, cardiac surgery is about as good as it gets. In many other clinical fields, key outcome data are simply not recorded, either because it has never been the practice to record such data or because such data are inaccessible. It is difficult to quantify the full extent of this, but at anecdotal level, based on the clinical research experience of the present author, it appears to be endemic. A stark example serves to illustrate this.

One area where audit is rarely carried out concerns monitoring the long term progression of patients with chronic disease, of whom there are many. The author has carried out research in this field part of which has involved helping to establish data collection procedures at a cardiology department to monitor the long term progression of all coronary artery disease patients attending outpatient clinics². Typically, such patients are followed up over many years. Such monitoring allows comparisons to be carried out of the clinical management patterns of different clinicians and their outcomes in terms of long term control of symptoms. The feasibility and usefulness of long term monitoring of the chronically ill is not surprising in itself, (although it is still rarely done). What came as something of a shock during this research exercise was the impossibility of extending the system to include records of death without a major data linkage exercise using information not readily available to the cardiac physician.

Unless, by good fortune, someone thinks to inform the cardiac physician that a out-patient attender has died, then there is no mechanism for this information to be added to audit records. The implications of this are profound when considering the topic of monitoring clinical outcomes.

Many, possibly the majority of secondary care clinicians responsible for the long term care of patients with chronic disease do not know with any degree of accuracy which or how many of their patients have died, let alone what was their cause of death. Thus rheumatologists do not in general know precisely how many of their patients have died from drug side effects

and cardiologists do not know how many of their patients on anticoagulation therapy have died from a stroke. The author has little experience of research in other fields involving chronic disease, such as diabetes, respiratory disease and mental health, but would expect the same pattern of lack of basic information.

Similar lack of basic information occurs in general practice, particularly in cities. A GP may have very good records of the patients he has seen but, knows very little about the patients he has not seen. Patients may move, leave the country or die and unless the GP is informed of this, he will not know. It is thus difficult even to derive an accurate estimate of a GP's list size and even for the most clear cut outcome measure, death, GP records may be inaccurate.

In terms of the widget-factory view of health care in the NHS, primary care and secondary care of the chronically ill represent a substantial part of the overall process, certainly dwarfing paediatric cardiac surgery in terms of the number of patients and treatments. However, for both, there is a huge paucity of basic information even to assess such a basic outcome measure as death. Recommendations concerning clinical monitoring should be based on an appreciation that in many aspects of health care, basic information is simply not available to carry out such monitoring.

5. WHAT OUTCOME MEASURES SHOULD BE USED?

As the evidence presented to the Inquiry has shown, it is surprisingly difficult to assess overall clinical performance even in the case of a very clear cut outcome measure such as death within 30 days of an operation. In other clinical fields, difficulties in assessing performance may be compounded, not least because death may not be a pertinent outcome measure.

For example, in orthopaedics, death is not usually the major issue whereas something like post-operative joint failure is. The number of potential outcome measures that could in principle be monitored related to clinical activity is enormous, a small arbitrary selection of examples serves to illustrate the variety and scope for monitoring performance:

- post-operative wound infection following gut surgery;
- rates of sexual dysfunction following prostate surgery;

- early recurrence of breast cancer following lumpectomy;
- rates of vaginal tearing during forceps delivery;
- mis-diagnosis or inappropriately equivocal diagnosis in pathology;
- delays and errors in the pharmacy process;
- delays in GP referrals to cancer assessment clinics;
- inappropriate referral rates by GPs;
- late or cancelled visits by health visitors;
- patient satisfaction with hare-lip reconstruction;
- dental filling replacement rates;
- inappropriate drug prescription rates;
- ineffective treatment for backpain.

Every speciality can be expected to have different requirements in terms of quality assurance monitoring. Even to list all the potential outcomes that might be considered would be a major exercise. It is almost certainly infeasible for hospitals to monitor all the outcome measures that could in principle be measured, thus there is the question of determining which should be ignored.

It may well be that there is a basic minimum set of outcome measures that all hospitals should reasonably be expected to monitor, or that should be monitored centrally. However, it is not clear what this set should comprise. Indeed, the determination of what should be in this minimal data set would in itself be a major exercise.

The possibility also needs to be considered that choosing to monitor a particular event may intrinsically effect the rate at which it occurs. Monitoring the numbers of patients whose postoperative hospital stay exceeds 7 days, as a surrogate measure of poor surgical outcome, could possibly result in many such cases being discharged prematurely.

Even with apparently clear cut outcome measures, such as mortality within 30 days of an operation, clinical practice may change to inflate apparent performance. Anecdotally, experience in the USA was that when published mortality tables were introduced concerning individual cardiac surgery centres, many centres reacted by adjusting their practice, avoiding the more difficult cases thus artificially improving their mortality ratings.

6. STATISTICAL METHODS TO ASSIST THE MONITORING OF OUTCOME

Medical statistics is a very broad subject and a wide range of statistical techniques have been developed for the analysis of clinical data. It is inappropriate in a brief report such as this to attempt to summarise all the techniques at the disposal of the medical statistician, however, it is useful to discuss some of the major issues that arise.

6.1 Observational versus non-observational data

The vast majority of data concerning NHS operation is collected at the time that health care processes are carried out and broadly speaking, serve to give a record of current patient status, what was done, when and by whom. In statistical terminology, such data are referred to as observational, being relatively unstructured observations of the health care process as it is being carried out. Observational data is very different in scope, and usually in quality, from the sort of data that would be expected for formal scientific projects which depend upon specifically designed and controlled studies, with regular, complete and uniform data collection protocols, quality control checks and pre-specified analysis plans.

The purist might disagree with the use of formal statistical methods, particularly with statistical significance testing, in the context of observational data. Statistical significance testing is more usually associated with the formal testing of hypotheses and tacitly presupposes scientifically well designed and controlled studies.

While the observational data sources available within the NHS are undoubtedly useful, it could not be claimed that the quality of routine data collection procedures within the NHS matches that which would be expected in formal scientific studies. As a result, credibility problems inevitably arise concerning the analysis and interpretation of such data.

In a Platonically ideal world, data collection procedures relating to the monitoring of clinical performance would be prospective, would follow formal protocols and would be carried out by individuals who are independent of the care processes being administered. Such monitoring should ideally be carried out with blinding, so that the assessor would not be aware of the identity of the clinician, the clinical centre or indeed the form of treatment

administered. Indeed there might be some that would advocate the gold standards of randomised controlled trial methodology.

In reality, the ideals of scientifically perfect non-observational data collection studies are unlikely to be feasible in all but a few NHS contexts. Indeed it is hard to envisage what the exceptions might be, outside the realms of research. The reality is that it is very expensive to carry out formal prospective clinical studies. While it is hard to generalise, formal trial-based studies carried out for research purposes usually cost in excess of £75,000- £100,000, often well in excess of such figures. The idea that a hard pressed NHS Trust could afford such costs for each of the units that were deemed necessary to monitor seems unrealistic.

Again, appealing to the widget factory metaphor of NHS operation, it is instructive to note that outside the confines of health care, quality control and process monitoring in manufacturing industry is usually based on observational data and rarely makes use of purist randomised controlled trial principles.

6.2 Issues associated with the analysis of observational data

Perhaps because observational data do not stem from carefully designed scientific protocols, there is a tendency, within medical statistics, to regard such data as being of questionable worth. One is reminded of Damon Runyon character who when asked why he continued to take part in a crooked craps game replied, "the game may be crooked, but it's the only game in town". This is very much the issue. Imperfect as observational data are, they are usually all that is available and it would be silly to ignore their usefulness.

That said, a large degree of common sense is needed in interpreting such data since they do not conform to the standards of formal scientific trials. It is often unwise to place undue reliance on formal statistical methods, such as hypothesis testing, in the analysis of observational data. To do so is almost to pretend that the data do stem from a formal controlled study. Statistical tools such as confidence and prediction intervals do serve a function in assisting the interpretation of observational data, but should not be regarded as having the same import or scientific meaning as in, say, the report from a rigorous pharmaceutical trial.

The author's experience of observational data is that they are usually very complex, usually have missing values or ambiguities, and do not follow uniform data recording procedures. This can make it very difficult to apply standard statistical methods directly and alternative, sometimes rather technical analysis methods are required¹³.

Given the complexity and general messiness of observational data, the principal aim in the analysis of such data is usually to distil a simple description of the overall features of the data. It helps if this is done in graphical form, although this may not always be possible.

Analysis of observational data thus commonly relies on finding a suitable means for distilling the overall essence of what is in a complex data set. There are a range of descriptive statistical methods that can be used to assist this process, many of which are mundane in technical statistical terms, but which provide a powerful armoury for summarising data.

A difficulty that needs to be acknowledged is that there isn't a systematic methodology for determining the most appropriate methods for summarising particular sets of observational data. The author's experience is that each new project brings new challenges and commonly, one needs to think up something from scratch.

If more widespread monitoring of clinical performance is to be introduced, and if this is for the most part to be based on the analysis of observational data, then there will be a need for considerable amount of further research to develop appropriate analysis methods for such data.

6.3 Analysis of observational data concerning adverse outcomes - Cusum methods

So called Cusum methods, introduced to the surgical community in a notable paper by de Leval et al⁹, and extended by others^{4,5,6}, seem to be the most popular means of examining overall performance in relation to cardiac peri-operative mortality and the analysis of other adverse events. Such methods are described in detail in another report prepared for the inquiry¹⁰. Here, only brief details will be given.

At their simplest, Cusum methods are based on charts showing a cumulative running total of total mortality plotted against total number of operations performed. This gives a jagged curve that climbs upwards. Changes in performance can in principle be detected by observing how the slope of the Cusum curve changes. Also, the performance of different clinicians can be compared by comparing their respective Cusum charts.

It is recognised that there are circumstances where simple Cusum curves may mask important features of clinical performance. If the case mix of a surgeon changes, which is likely for a surgeon in training say, then apparent stability of the Cusum chart may shield the fact that the surgeon is dealing with more and more difficult cases, while becoming more and more skilled. Equally, an apparently alarming rise in the slope of an experienced surgeon's Cusum curve may indicate nothing more than he has taken over the cases of the Professor, during a period of absence, and thus his case load has acquired a higher intrinsic risk.

More sophisticated analysis is available in the case where pre-operative risk factors are known, and estimates are available for the risks of death, which are possibly different for every patient. Methods for forecasting peri-operative death risks are available in the case of adult cardiac surgery¹¹. Using these, one can take the surgeon's case mix into account and construct charts of overall performance based on 'Net Life Gain', compared with the mortality that would have been expected given the case mix^{4,5,6}.

Such methods are largely descriptive and do not attempt to represent formal statistical hypothesis testing. Alternative methods for examining the statistical consequences of sequences of adverse events are currently the subject of research. New methods are emerging¹⁴ and more can be expected.

A central issue in such analysis is the generation of estimates of the probability that an adverse event will occur based on factors that can be assessed prior to the clinical process under scrutiny. In the case of adult cardiac surgery fairly good estimates of operative risk are available in terms of factors such as a patient's age and cardiac status^{11,5}. As pointed out by Stark¹ in the case of paediatric cardiac surgery, modelling case mix risks is far less advanced. As far as the author knows, relatively little research has been done concerning factors

affecting adverse event rates in broader clinical fields, although this may be due to ignorance of the relevant literature.

If broader based systems are to be introduced for monitoring adverse event rates, then research into case-mix factors and their relationship to adverse event rates is a high research priority.

6.4 The arbitrary nature of statistical significance testing

Although observational data is likely to be the prime source of information for clinical monitoring, one should not ignore alternative, more formal scientific studies that might be considered. Nor indeed the potential for using formal statistical methods in relation to the analysis of observational data. In this respect, a note of caution needs to be given about the interpretation of such analysis.

It is common in clinical studies to carry out statistical hypothesis testing. For example, if the birth weights of babies were measured for 50 mothers who were smokers and 50 mothers who were non-smokers, then differences in birth weight between the two groups might be described as 'statistically significant at the 95% level'. This means that if birth weights were not systematically different between the two groups, then the difference actually observed would be very unlikely, indeed the estimate of its occurrence would be judged to be 1 in 20 (5%) or less.

The use of 95% rather than 85% or 97% is purely a matter of custom and practice, indeed some studies use tests based on 99% instead. Although the higher the figure, the more certainty there is that the conclusions are sound, it should be recognised that these values are largely arbitrary. In spite of this, 'statistical significance' has become something of a mantra within clinical research. If a study does not show statistical significance, it is often regarded as having failed, to show no evidence that differences exist, or even worse, that this represents strong evidence that there is no difference. This is very dangerous and misleading use of statistics.

A surgeon should not be complaisant that his performance is adequate purely because his performance curve persistently hovers on the righteous side of the 95% threshold. Equally, a surgeon whose loss of 5 consecutive patients within 30 days of operation pushes him beyond the arbitrary 95% limit, should not necessarily be demonized. Even if there were no mitigating external factors, this could just be a 1 in 20 coincidence.

In terms of performance monitoring, statistical testing can only ever be a guide.

7. WHAT CONSTITUTES ACCEPTABLE PERFORMANCE?

As discussed in the previous section, there are statistical methods that provide a means for distilling information from complex clinical data and which can help clinicians and managers to develop a clearer view of overall performance and assist with the task of alerting clinicians and managers to potential problems. No doubt, additional methods will become available as further research is done. However, what cannot be expected to emerge purely from statistical research is arbitration about what constitutes acceptable performance and, equally important, what constitutes acceptable divergence in performance between centres and between clinicians. This goes way beyond the purely technical issues of what statistical tests should be used and what level of significance should be adopted.

Acceptability of treatment quality may well depend on what the treatment is trying to achieve and what outcome measures are being used to for assessment. Presumably, one would be less concerned about the number of patients reporting mild pain during a smear test, than about overall mis-diagnosis rates, although both are important.

It is intrinsically complex to assess acceptability in rational terms. The odd mistake in the treatment of respiratory infections may appear dwarfed in importance compared to paediatric deaths following surgery, yet there are far more deaths due to respiratory disease than to congenital heart defects.

A commonly held view of acceptable performance, is that it reflects what is typical in practice. This is implicitly the method that has been used to assess performance at Bristol, by

comparing its mortality to mortality aggregated over several comparator sites⁷. While this seems sound for cardiac surgery, there are other clinical contexts where merely observing what happens at centres in the UK does not give a particularly useful view of what is acceptable. For example, is it acceptable performance for the visually impaired to have to wait several months for a cataract operation? It may well be common practice, but that is not to say that this is the standard that should be strived for.

What constitutes acceptable performance is thus a very difficult issue, and it is by no means clear who should be the arbiters of what is acceptable and what is not. Should this be clinicians, the Royal Colleges, policy makers or indeed the general public?

8. INHERENT DIFFICULTIES IN PURE RELIANCE ON STATISTICAL METHODS

Important issues arise concerning monitoring pointed out by Frankel et al¹² in relation to mortality variations between general practitioners' practices. In the wake of the Shipman case, it had been suggested that there might be formal monitoring of practices to detect divergence in mortality.

Crude mortality rates are approximately 1.1% per annum. Based on this, and a given practice size, routine statistical methods allow one to estimate the expected annual number of deaths in a practice. One can also estimate an upper threshold for the annual mortality within a practice above which the mortality rate would be judged to be anomalously high according to 99% confidence intervals.

If such a system were to be introduced to trigger further inquiries into potentially aberrant practice, then 0.5% of practices would be expected to have an annual mortality above the crucial threshold value. With some 9000 practices in England, this would mean that annually, some 45 practices would be deemed to warrant further investigation.

Further, for Shipman's practice of 3600 patients, the expected annual death rate was 40 and the threshold value that would trigger suspicion, based on 99% confidence intervals, is 58.

Thus, quoting Frankel et al¹² *'Shipman's practice of 3600 would allow an excess of 18 deaths a year above average to pass as unremarkable, which is more than the 15 murders over three years he is currently convicted of...'*

This elegantly illustrates the folly of over-reliance on statistical methods in relation to clinical monitoring. The more clinical processes that are subjected to formal monitoring, the more that will appear divergent, by chance coincidence. Equally, truly aberrant performance may go undetected if statistical methods are the only methods used for monitoring.

9 HOW MUCH DATA SHOULD BE COLLECTED AND BY WHOM?

9.1 Existing national data sources

A clear message emerging from the statistical evidence commissioned by the Inquiry is that, for the purposes of assessing overall performance monitoring in cardiac surgery, there are problems in placing overmuch reliance on the two major data sources, Hospital Episode Statistics (HES) and the Cardiac Surgery Register (CSR). It should be noted that part of the statistical work commissioned by the Inquiry was an exercise to scrutinise the data collection processes and data quality of the CSR⁸. Unfortunately, no similar exercise was commissioned concerning the data quality of the HES, although there is some evidence of substantial mis-coding and incompleteness. Crucially, there were some major discrepancies between HES and CSR concerning estimates of case load and numbers of deaths for the 12 categories of surgical procedure, aggregated over all of the comparator sites. It is difficult to believe that HES is close to perfection and that all this divergence is due to errors in CSR. Indeed, due to the diagnostic coding scheme used within HES and its non-clinical use, cardiac surgeons have more confidence in CSR. Quoting from the final Report on the Overview of the Statistical Evidence⁷

" [From 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 report points out, is only capable of detecting gross divergence. 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 author does not agree with the view that data collection systems can be rectified by relatively minor modifications, even for the purpose of monitoring outcomes in cardiac surgery. When considering clinical monitoring in broader areas of the health sector, particularly when outcome measures are not death, the author would expect HES to be a wholly inadequate basis for clinical monitoring.

Consideration should be given to whether the whole process of national data collection should be thoroughly reviewed and perhaps reorganised.

9.2 More general data for monitoring clinical performance

As discussed by Stark¹, data requirements for effective monitoring of performance in paediatric cardiac surgery go considerably further than HES or CSR, and if future monitoring is to be carried out in relation to outcomes of such surgery, the data collection processes envisaged by Stark, or something close to them, are likely to be required. However, Stark focuses purely on paediatric cardiac surgery. While many of his observations are generally relevant to many clinical fields, the organisation of data collection in other clinical specialities and for other outcomes and performance measure, is likely to be have markedly different requirements. At a stage where one is not even sure what outcomes and performance measures are appropriate, nor which clinical processes should be monitored, it is perhaps folly to be too specific about data collection and analysis needs.

10. IS MONITORING OF CLINICAL PERFORMANCE ALWAYS DESIRABLE?

The commissioning brief for this report is very much concerned with asking advice on ways for improving statistical aspects of clinical monitoring. Almost implicit in this is the view that accurate performance monitoring and analysis is inherently a good thing. As with all views that seem self-evidently true, one needs to reflect on whether this is actually the case.

10.1 Interaction between monitoring and other health care processes

As the emergency resuscitation example discussed in Section 3.3 illustrates, there are circumstances where it is by no means clear that monitoring is feasible or desirable. It is clearly more important for the crash team to get on with their life-saving efforts rather than worry about detailed and accurate record keeping.

This is not an isolated example and there are many cases where the processes of health care need to be balanced against the needs for collecting accurate data. Another, much more frequent occurrence, concerns GP consultations. On average these take a surprisingly short

length of time, of the order of a few minutes. If, in addition to the normal things that go on in a consultation, GPs also had to record additional data that could be used to monitor their performance, this would extend the consultation process considerably. The author's personal experience is that even the time taken to input and process information about drug prescriptions represents a substantial part of the consultation process. Additional data collection could easily extend the consultation process by 50%-100%. Before recommending monitoring in the field of General Practice, there is a need to be considered whether it is practical or, indeed, ultimately beneficial.

In the case of the crash team, provision of an additional team member whose job is to monitor the timing of events, could in principle allow accurate monitoring to take place, although at inordinate cost. In the case of the GP, the constraint is more concerned with time subtracted from other consultation matters. In both cases, while monitoring might provide useful information, it conflicts with effective health care.

The widget factory metaphor for health care is useful here. What is important is the overall effectiveness of the system. The relevance and usefulness of monitoring needs to be judged in terms of the performance of the overall system, not purely as an end in its own right. If elaborate and costly quality control in one part of the widget factory has little or no beneficial effect on the performance of the system as a whole, or if disrupts production unduly, it should not be introduced.

10.2 Should clinical monitoring be based on outcomes or process?

The nature of the Inquiry has inevitably coloured views so that it naturally seems reasonable that performance monitoring to be concerned with assessment of outcome, since this has been a principal concern of much of the evidence. Yet it is sensible to consider whether outcome is really the key issue, or whether there are circumstances where it is more important to assess what it is that clinicians do, rather than what is the outcome.

Evoking the widget factory metaphor, a foreman wouldn't consider mounting a randomised controlled trial of productivity if he found a worker whose view was that best butter was the ideal lubricant for a high speed lathe.

Rather than monitoring outcomes, perhaps it is preferable to examine what clinicians actually do. In regard to this, the lay view is probably that there is very little variation from one clinician to the next in terms of the clinical processes that will be carried out for a patient suffering from a given complaint. The reality is very different. Large differences in clinical practice are common.

Before recommending intricate data collection and analyses of outcome, the Inquiry panel should consider whether it might be more productive to focus on variations in clinical practice rather than outcome monitoring.

10.3 Health economic consequences of performance monitoring

As the crash team and GP examples illustrate, attempts to improve standards of monitoring may conflict with other factors that are intrinsic to the provision of effective health care. Needless to say, a major factor to consider in this light is the cost consequence of performance monitoring. Quoting from the executive summary the Stark's report on monitoring in paediatric surgery¹:

"None of the above (recommendations) are achievable without adequate, separate funding for equipment, infrastructure, personnel and training..."

This is undoubtedly a sensible conclusion, although whether such additional funding will be made available is a matter for others to decide. It is inevitable that if monitoring of clinical performance is judged to be essential, then it needs to be funded. It will not happen for free. The reality of NHS operation is that staff already work to capacity; additional tasks related to performance monitoring cannot be taken on unless time spent on other tasks is reduced.

Stark's report¹ does not include cost estimates, but such costs are likely to be substantial, even considering paediatric cardiac surgery in isolation. One can only speculate about the health economic consequences of wide scale performance monitoring within the whole of the health sector, however one might guess that if introduced, it could become a major item of NHS expenditure. The knock on effect of this would be to reduce the funding available for health care delivery. Pragmatism suggests that great care should be taken before such a

radical change in NHS operation is recommended and much further evidence is needed that introducing widespread performance monitoring does actually result in improved overall standards of health care delivery.

11. REFERENCES

1. J. Stark, 'Future improvements in the routine monitoring of surgical performance', Bristol Royal Infirmary Inquiry, 2000.
2. C. Sherlaw-Johnson, J. Mitchard, S. Gallivan, D.L.H. Patterson, T. Treasure (1995) 'Displaying the Long Term Progression of Patients with Coronary Artery', British Heart Journal; **74**:559-562, 1995
3. J. Stark, S. Gallivan, J. Lovegrove, J.R.L. Hamilton, J.L. Monro, J.C.S. Pollock, K.G. Watterson, 'Mortality rates after surgery for congenital heart defects in children and surgeons' performance', Lancet **355**:1004-1007, 2000
4. J. Lovegrove, O. Valencia, T. Treasure, C. Sherlaw-Johnson, S. Gallivan, 'Monitoring the result of cardiac surgery by variable life adjusted display (VLAD)' Lancet; **350**:1128-1130
5. J. Lovegrove, C. Sherlaw-Johnson, S. Gallivan, 'Monitoring the performance of cardiac surgeons' Journal of the Operational Research Society; 50 (Number 7): 684-689, 1998.
6. C. Sherlaw-Johnson, J. Lovegrove, T. Treasure, S. Gallivan, 'Likely variations in perioperative mortality associated with cardiac surgery: when does high mortality reflect bad practice?' Heart **84**:79-82, 2000
7. D.J.Spiegelhalter, S. Evans, P. Aylin, G. Murray, 'Overview of statistical evidence presented to the Bristol Royal Infirmary Inquiry concerning the nature and outcomes of paediatric cardiac surgical services relative to other specialist centres from 1984 to 1995', Bristol Royal Infirmary Inquiry, 2000.
8. A.E.Lawrence and G.D.Murray, 'The UK Cardiac Surgical Register: assessment of data quality issues for the Bristol Royal Infirmary Inquiry', Bristol Royal Infirmary Inquiry, 2000.

9. M.R. de Leval, K. Francois, C. Bull, W. Brawn, D.J. Spiegelhalter. 'Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations' *The Journal of Thoracic and Cardiovascular Surgery*, 107(3): 914-924, 1994.
10. S. Gallivan, 'Learning curves in relation to surgery', *Bristol Royal Infirmary Inquiry*, 2000
11. V. Parsonnet, D. Dean, A.D. Berstein, 'A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease.', *Circulation (supplement I)* ; 779(6): I-3 - I-12, 1989.
12. S. Frankel, J. Sterne, G.D. Smith, 'Mortality variation as a measure of general practitioner performance: implications of the Shipman case', *British Medical Journal*; **320**, 489, 2000
13. C. Sherlaw-Johnson, S. Gallivan, J. Burridge, 'Estimating a Markov transition matrix from observational data.', *Journal of the Operational Research Society*, **146**:405-410,1995
14. S. Steiner, R. Cook, V. Farewell, 'Monitoring paired binary surgical outcomes using cumulative sum charts' *Statistics in Medicine*, 18:69-84, 1999