

Monitoring Clinical Performance: a Statistical Perspective.¹

**David Spiegelhalter, Gordon Murray, Klim McPherson,
Alison Macfarlane, Stephen Evans, Robert Curnow,
Michael Campbell, Paul Aylin.**

November 2000

¹ ©Crown Copyright 2000. 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 authors alone, and do not necessarily represent the views of the Bristol Royal Infirmary Inquiry.

Executive Summary

1. Our experience as experts and analysts for the Bristol Royal Infirmary Inquiry convinced us that the current 'dual' data collection in separate administrative and clinical systems is wasteful and anachronistic.
2. Performance monitoring should form a component of a sustained quality improvement programme - there is little value in measuring variability in performance without a rigorous subsequent search for explanation and action.
3. Performance monitoring comprises an evaluation model supplied by a suitable information system. An evaluation model concerned solely with short-term surgical outcomes could use the type of system described by Stark [1], but this 'card-index' paradigm has limited scope.
4. An integrated 'population-based' evaluation model, using linked data sources, could monitor the entire process of care of an individual and form an evidence-base for systematic quality improvement. HES cannot currently fulfil the information requirements for such a model. Clinical record systems exist that may fulfil this role as part of routine care, but substantial resources may be necessary.
5. All monitoring systems have a cost, both in terms of resources and potential side-effects, which needs to be balanced against possible benefit. US experience of the impact of 'report-cards' should be carefully studied.
6. In our view, a properly monitored Cardiac Surgical Register could have drawn attention to Bristol at an early stage, and an improved version may be adequate for such a role in the future.
7. Current and proposed schemes within clinical governance and other initiatives raise subtle quantitative issues. Skilled statistical input is needed in both design and analysis of performance monitoring, just as in controlled trials and epidemiological studies.
8. A vital aim should be two-way linkage between administrative and clinical systems. Protocols for mapping reduce the need for an enforced common coding scheme.

9. A large amount of data may be required to confidently identify divergence in performance, and it is important that unrealistic expectations should not be created.
10. Appropriate statistical techniques for surgical monitoring include confidence intervals for risk-adjusted rates, and cumulative sums of risk-adjusted outcomes.
11. When a large number of comparisons are made, there is a danger of excessive false-positives arising from the naive use of significance tests or, equivalently, confidence intervals excluding thresholds. Existing statistical techniques for dealing with this issue need to be explored.
12. The practice of ranking institutions to create 'league tables' is inappropriate, since ranks are notoriously sensitive to chance variability. If necessary, techniques exist for placing confidence intervals around observed ranks.
13. Institutions that have been identified as 'extreme' tend to become less so when re-examined ('regression to the mean'), since part of the reason for their extremeness was a run of good or bad luck. This simple phenomenon could lead to spurious claims being made about the benefit of interventions to 'rescue' failing institutions.
14. As high-quality integrated databases become available, joint clinical-epidemiological investigation of causes for variability will require careful use of statistical methodology, in a culture that rigorously seeks explanations rather than ascribing 'blame'.

1. Introduction

This report was commissioned by the Bristol Royal Infirmary Inquiry to assist the Panel's consideration of key directions for the future in monitoring clinical performance. The Inquiry brief was to provide a relatively brief report of high-level expert advice to the Inquiry Panel which would aim to distil - on the basis of personal statistical or epidemiological expertise and judgement - the potential contribution of statistical and/or epidemiological perspectives, methodologies or techniques in achieving more effective monitoring of clinical performance in future.

We begin in Section 2 with some general principles of monitoring systems, emphasising the contrast between a *surgical* model concerned solely with short-term surgical outcomes, and a *population-based* model with the potential to examine a child's entire process of care. Section 3 describes the potential role of statistics in performance monitoring systems and attempts to relate this to current practice and proposals, while Section 4 steps through the detailed components of a monitoring system, noting the statistical aspects and making recommendations as to future practice.

Our remarks are limited to our own professional domain and so do not, for example, concern the organisational aspects of such systems, such as the body with responsibility for monitoring and investigating possibly divergent performance. Neither do we provide detailed critiques of current practice or proposals, although we do try to take account of recent initiatives in the NHS and elsewhere. We do, however, include data collection within our remit as good data are a pre-requisite for reliable statistical analysis. In our work as statistical experts and analysts for the BRI Inquiry, judgements had to be made on the basis of multiple sources of data of poor or at best mediocre quality: in particular, we were faced with 'dual' data collection in which administrative and clinical recording systems acted in parallel and with negligible communication. We understand why this situation has arisen in a climate with little motivation to collaborate, and both paradigms have worthy historical antecedents in, for example, Nightingale and Codman [2]. However, to an outsider this division now appears somewhat absurd and increasingly wasteful and

anachronistic, and a running theme of this report is the need for a single reliable source for each item of data, which may then be shared between linked systems. We hope our analysis will be useful in informing the Inquiry's wider consideration of key directions for the future.

2. Background

2.1 Why do we want to monitor clinical performance?

Knowledge of clinical performance can be valuable to a number of groups [1].

Somewhat simplistically:

- *Patients* and their carers should, ideally, be provided with appropriate individual risks of short-term mortality, as well as long-term morbidity and survival prospects.
- *Clinicians* should be aware of their individual performance over a sequence of cases in order, for example, to compare with professional standards.
- *Managers and clinicians* at a local level need to be aware of comparative performance of both individual clinicians and teams.
- *Commissioners of services and public health departments* at a local level wish to monitor the health of specified populations.
- *Policy-makers and researchers* at a national level should be able to make strategic decisions based on firm evidence.
- The *clinical-scientific community* seeks understanding of the underlying causes of variability in outcome, and hence how to improve future performance.

While some of these groups may have similar requirements, a single monitoring system is unlikely to satisfy all parties.

2.2 Archetypal models for monitoring performance

It is important to distinguish the *evaluation model* which specifies the performance being monitored, and the *information system* from which the necessary data are

derived. The former should ideally drive the latter. The range of requirements noted previously lead us to delineate two broad models for performance comparison, which we shall term surgical and population-based.

Surgical evaluation model. Stark [1] provides a framework that focuses on short-term mortality, and would allow scrutiny of results following introduction of new techniques, surgical learning and so on. The corresponding information system would be a simple stand-alone database with standardised data collection and trained medical input.

Critique: Such a model may be appropriate for monitoring professional surgical standards, but does not address other possible objectives. The information system follows a 'card-index' paradigm, with no provision for linkage to local or national administrative data-bases, and hence longer-term morbidity and mortality prospects for patients cannot be assessed. In addition, the databases only contain patients who came to surgery and hence cannot be used to assess outcomes of all babies with cardiac problems, in particular the proportion dying before surgery. Hunter et al [3] stress that 'the quality of management and the outcomes of treatment must be considered in relation to all infants born with heart disease', and that the minority of such children requiring surgery represent 'the tip of the iceberg'.

Population-based evaluation model. This would examine the way in which a health-care system has an impact on outcomes for a defined population. Ideally, this would involve monitoring of the way the condition is detected and registered, referral to secondary and possibly tertiary care, decisions about treatment, and short and long-term outcomes in terms of mortality and, most importantly, morbidity. The information system for such a model should ideally be an integral part of routine care. When investigating the outcomes associated with specific components of the system, such as a unit or an individual surgeon, data should then be available concerning pre-existing health status and risk factors.

This model is child-centred rather than procedure-centred and, by linking episodes of care, analyses should be possible across professional boundaries.

Rigorous epidemiological studies using controls become feasible, the need for which is being increasingly recognised within systems such as the Confidential Enquiry into Stillbirths and Deaths in Infancy (CESDI) [4].

Critique: Such proposals may seem optimistic, both in looking beyond short-term outcomes and in the information requirement. HES does not provide an adequate basis for such a model. However, serious attempts are being made to improve the situation through record linkage and through the establishment and improvement of disease and congenital anomaly registers [5]. In paediatric cardiology, systems such as *HeartSuite* are essentially electronic patient records and appear to have many of the desired characteristics [6]. There is, of course, a danger in embarking on such an ambitious exercise and collecting too much data without a clear idea of what is to be done with them. Progress towards this goal needs to be carefully staged.

These two models have many common features, are certainly not mutually exclusive and should ideally be complementary. Efficiency and accuracy point to careful linkage as being a pre-requisite of any system.

2.3 Performance monitoring and quality assessment

Attention is currently strongly focussed on early detection of divergent surgical performance, but this forms only a part of the process of care received by an individual. It can be argued, from both clinical and epidemiological points of view, that a programme of sustained quality improvement requires deeper understanding of the mechanisms that influence outcome in routine care - there seems little value in measuring variability in outcomes unless explanations can be found and appropriate action taken [7]. The limited role of outcome measurement in a quality improvement programme has long been realised in industry. Monitoring clinical performance is thus ideally a by-product of routine care, and done in tandem with evaluations of the risks and benefits of different forms of care through clinical trials and other methods, and in the light of knowledge of the characteristics of populations for whom clinicians are caring. Experience with the BRI Inquiry suggests that such a broader, 'systems', perspective is desirable, provided it is implemented in a rigorous scientific manner.

The cost of introduction of a distinct system of performance indicators must be carefully weighed against the potential benefits. This cost not only covers monetary and staff resources, but also staff morale and the possibility of perverse incentives arising through over-focus on specific indicators. There has been extensive study in the US on the potentially negative impact of the introduction of 'report-cards', e.g. [8]. It should not be taken for granted that a sophisticated, expensive system is necessary: the apparent excess mortality in Bristol could have been identified at an early stage by the somewhat crude existing system, had there only been an adequate provision for analysis, feedback and action [9]. For the limited purpose of monitoring surgical mortality, an improved UK Cardiac Surgical Register might well be sufficient to provide early warnings of sufficiently large divergent performance.

3. The statistical perspective

3.1 The role of statistics in clinical performance monitoring

Any system for monitoring clinical performance can be broken down into the following components: 1) data selection, 2) data collection, 3) data summarisation, 4) setting benchmarks, 5) comparison with benchmarks and subsequent action, 6) communication of risks and comparative performance, and 7) investigation of causes of variability. Our experience with the BRI Inquiry and other studies suggests that statistical thinking has a role to play in each of these stages, particularly 5 to 7. It should be emphasised, of course, that statistics is only one component of monitoring clinical performance.

3.2 Potential role of statistics in clinical performance monitoring

Relevant initiatives include the following.

- The *NHS Information Strategy* [10] and *NHS Plan* [11] envisage growing use of electronic patient records, linked through the NHS number so that information is entered only once, allowing analyses based on long-term follow-up, and

developing and improving congenital anomaly registers and linking data to births and deaths at a national level.

- *National Service Frameworks* are being established, with core datasets and audit packages, which will require formal techniques for identifying deviations from specified levels of service. The National Institute for Clinical Excellence (NICE) and the Commission for Health Improvement (CHI), as well as the Clinical Standards and Health Technology Boards for Scotland, are engaged in setting guidelines and monitoring process measures to assess adherence to those guidelines.
- An *Organisation with a Memory* [12] recommends formal collection and analysis of data on all adverse events.
- *NHS Performance Indicators* [13] based on routine administrative data make use of indirect standardisation, confidence intervals, and explicitly rank institutions and authorities according to outcome measures.
- The *Scottish Clinical Outcomes Indicators* [14] display trends in outcomes for institutions and authorities, with some risk adjustment and 95% intervals.
- The *Cardiac Surgical Register* has recently started a programme of systematic analysis and publication of their results [15].
- The *Central Cardiac Audit Database (CCAD)* [16] appears to be attempting a population-based approach: clearly a lot of attention has been focussed on issues of information entry and transfer, although we are unaware of future proposals for analysis.

Such projects are, with certain exceptions, seen primarily as the domain of clinicians and administrators, and there is minimal statistical involvement beyond, say, calculation of confidence intervals. The subtlety of the quantitative issues suggests that skilled statistical input is needed in both design and analysis of performance monitoring, just as in controlled trials and epidemiological studies.

3.3 Some general statistical principles

Before examining the components of a performance system in detail, we list some statistical principles which showed their relevance within the BRI Inquiry.

- In making comparisons, we may seek to increase precision but not at the expense of bias, since we know how to quantify random variability but cannot compensate for systematic bias.
- It is *better to get an approximate answer to the question you asked, rather than a precise answer to a different one*. Both objectives are served by standardised data collection.
- Although statistical techniques are designed to allow for the effect of random variability, there is a danger of too rigid use of significance tests, or equivalently 95% confidence intervals.
- We can quantify the ability to detect differences in performance, which is particularly relevant with low event rates. For example, the current mortality for open cardiac surgery on under 1's is around 7.5% [15]. Suppose a centre's true mortality were twice the national average, i.e. 15%, then to have an 80% chance of detecting this (with 95% confidence), we would need to monitor 120 operations, which may require several years to accumulate in a smaller centre. The ability to detect such important differences is driven by the number of adverse events, rather than the total activity. The number of adverse events being monitored may be increased, either by examining trends over a number of years, or by including 'near-misses' or serious morbidity as adverse events. It may therefore be not at all easy to detect even large differences in rates of adverse outcomes.
- Statistical principles need to be clear and comprehensible to all parties, although the precise computational details need not be.

In addition, it is worth emphasising some relevant principles in data collection:

- Any item of data should only be entered once [10].
- As much thought and effort needs to be put into getting information out, as is given to getting data in. This should include feedback to those supplying the data.
- A sense of ownership and responsibility encourages data accuracy.

4. Statistical input into monitoring systems

4.1 Data selection

Diagnoses and procedures: Stark [1] discusses the current lack of an agreed coding scheme and the multiple existing proposals. While this is indeed a drawback for detailed clinical work, it should be emphasised from a statistical perspective that only a limited number of categories will be used in any monitoring system, and hence provided a consistent mapping can be made to those categories the precise details of the coding system is not a vital issue.

Outcome measures: Accurate ascertainment of mortality is essential, and linkage should allow longer-term outcomes to be obtained. Definition and attention to both short-term 'near-misses' and longer-term morbidity both increases the sensitivity of a monitoring system by increasing the number of events, and satisfies the demands of proposal for monitoring clinical errors [12]. However, as soon as subjective measures are introduced, there is a possibility of bias or manipulation.

Risk factors: The surgical model emphasises the collection of risk factors just prior to surgery, while a systems model focuses on factors present when the child first came in contact with the service [9].

Process measures: Records of non-operative treatment may both help explain individual adverse events and can be investigated as explanatory factors for apparently divergent performance. Quality assurance in the US is increasingly focussing on process measures and adherence to guidelines, and the work of NICE and CHI, as well the Clinical Standards Board of Scotland, will lead to increased demands for such data in response to published guidelines.

This apparently points to the selection of large quantities of data, and this must be balanced against the burden of collection. We follow Stark [1] in emphasising that data items should only be included if they have a clear purpose.

4.2 Data Sources and Collection

Any performance monitoring system relies on accurate and complete data, and this requires conscientious staff, training in standardisation, and appropriate scrutiny. Current systems fall far short of this ideal, and an anomalous system has arisen whereby coders enter detailed clinical data which is subsequently distrusted and dismissed by clinicians. Stark [1] feels that only senior medical staff can be entrusted with data collection. However, any aspiration beyond the limited surgical model will require linkage to administrative data, following the growing trend towards linked medical information systems using the NHS number [11]. It should be noted that the Scottish clinical outcomes analysis [14] works from linked administrative systems.

It appears essential for there to be two-way linkage between administrative and clinical systems: clinical systems to download administrative and follow-up data, and administrative systems to download clinical data. Each should be able to map onto common diagnostic and procedural categories for monitoring. This is technically feasible [6] and surely should be the objective of any investment in systems. There should also be a facility for export of data for more complex analysis when the need arises.

4.3 Data summarisation

Before any formal statistical methodology can be used, it is necessary to specify basic quantities in order to derive appropriate summary statistics such as rates. These include:

Activity: this forms the ‘denominator’ in any rate. Paediatric cardiac surgery presents particular difficulties since a child may have multiple admissions, multiple operations per admission and multiple procedures per operation. A surgical model may define activity in terms of operations, but even this requires careful definition. A population-based model would consider the child as the ‘activity’. Furthermore, if estimating the rate at which events occur over time, the relevant ‘start-time’ needs careful specification.

Events: this forms the ‘numerator’ in any rate. While a surgical model might focus on short-term mortality such as deaths within 30 days, a population-based model will examine longer-term outcomes.

Aggregation: to obtain sufficient events there must be aggregation of activity into a limited list of well-defined categories. This will depend on the benchmarks being set. In addition there is a need to aggregate over time as well in order to have a sufficient number of adverse events for effective analysis: for example, three years activity may be necessary to obtain confident results.

Risk-adjustment: a risk-adjustment scheme quantifies the expected risk associated with each unit of activity in 'standard' circumstances, which can be traded off against any adverse events that occur. Schemes based on pre-operative risk factors are in their infancy: a population-based model would only adjust for factors entirely outside the influence of the health service, such as unchanging pathological anomalies.

4.4 Setting benchmarks

It is important to understand that benchmark performance is based on unobservable 'true' rates, and that whether an individual or unit has exceeded those benchmarks over a specified period is a matter of statistical analysis.

Choice of indicator: Only a limited list of (possibly risk-adjusted) rates can be considered. Extremes include (i) to consider all open operations on under and over 1's, or (ii) to examine outcomes at a very fine diagnostic or procedural level. Between these two extremes there are intermediate options: e.g. (a) to select specific well-defined benchmark operations, such as those recently adopted by the Cardiac Surgical Register [15], or (b) to map all operations onto a limited list of groups, as adopted in some US analyses [17].

As noted previously, inclusion of 'near-misses' increases the adverse event-rate and hence increases the sensitivity to divergent performance. However, this could have unintended consequences if they are process measures: for example, if a 'near miss' is defined as placing a child back on bypass [18], this might induce reluctance to put children back on bypass when it would be of value to do so.

Choice of threshold: Thresholds are either absolute or relative.

- *Relative* thresholds are defined according to current overall performance. For example, the New York adult cardiac system [19] system uses state-wide mortality rate as a benchmark. Taken to its logical conclusions, this approach seeks to penalise those who are simply below average which, unless everyone is identical, may well identify those with no serious divergence.
- *Absolute* thresholds rely on specification of the high end of ‘acceptable’ performance, and possibly the bottom end of ‘unacceptable’ performance with a grey area in between.

The crucial distinction between relative and absolute thresholds is that the latter could lead to everyone having acceptable performance, while the former is essentially certain to identify someone as being divergent .

4.5 Detecting divergence from benchmarks and taking action

These two aspects are considered together since statistical criteria cannot be separated from the action to be taken. We again emphasise that it is not our role to specify action and who should take it, but we do recommend that a skilled statistician should be involved. There is a need for care and flexibility in this delicate and crucial activity, and in particular we follow the industrial quality control literature in pointing to the possible need for ‘warning’ and ‘alarm’ thresholds to indicate two levels of action, such as ‘internal’ and ‘external’ scrutiny.

Relevant statistical issues include the following:

1. Graphical techniques for displaying data summaries are important. For example, when monitoring the performance of a single individual or institution over time, one can plot the cumulative excess of observed over expected mortality [20] [21], although care is required with assessing the statistical significance [22].
2. A popular method when comparing centres is to plot the observed performance and 95% confidence interval: see for example the NHS Performance Indicators [13]. If the interval does not overlap a benchmark then attention focuses on that centre. However, by chance alone one can expect 2.5% of centres to be so identified, even if they are actually performing at the benchmark level. This indicates the need for caution in interpreting ‘statistically significant’ results.

3. Risk-adjustment generates expected outcomes E which is then compared with observed outcomes O - this is 'indirect standardisation'. For example, if performance is measured by mortality, one can compute the standardised mortality ratio O/E or the mortality difference $O-E$: the latter has been termed 'excess mortality' [9], although an alternative term might be preferable. However, it would be misleading to claim that statistical procedures can ever fully adjust for pre-existing risk factors, and so unadjusted outcomes should also be provided.
4. As already described, there is considerable danger in generating spurious 'false-positive' findings when carrying out many comparisons. Statistical techniques exist for dealing with these, such as Bonferoni adjustments and shrinkage estimation [23], although these can lead to an excess of 'false-negatives' in which genuinely divergent behaviour goes undetected. Again, care and flexibility are required.
5. The practice of ranking institutions to create 'league tables' is inappropriate, since ranks are notoriously sensitive to chance variability. If necessary, techniques exist for placing 95% intervals around observed ranks [24].
6. '*Regression to the mean*' describes the tendency for institutions that have been identified as 'extreme' to become less extreme when monitored in the future - put simply, part of the reason for their extremeness was a run of bad luck. This simple phenomenon could lead to spurious claims being made about the benefit of interventions to 'rescue' failing institutions.

Perhaps the most important idea is that there is little gained from measuring variability in outcomes unless one can suggest underlying causes and remedial interventions. Someone must always be bottom of any 'league table', and the vital issue is whether they are truly divergent and, if so, why? Investigating the underlying reasons for variability in outcomes is not straightforward: while adjustment for case-mix is in principle possible, one must keep in mind that clinicians may respond to individual patients' situations in different but appropriate ways. Investigations must value that clinical skill.

4.6 Communication of risk and comparative performance

Patients and carers have a need for individualised risk assessment for their specific circumstances. When a short-term mortality risk is required, this requires judgement on the appropriate diagnostic category and risk-adjustment procedure, as well as a decision on whether to use local or national figures, or a compromise, and over what time period. Methods exist for ‘partial pooling’ of local with national figures [25], but their application requires further investigation.

Clinicians and institutions need to be made aware of their performance in the context of a ‘safety’ rather than a ‘blame culture’ [12].

4.7 Investigation of reasons for variability in performance

This vital area is in its infancy, although there is growing interest in the extent to which process influence outcomes. We forecast that, as the quality of data sources improve, these studies will follow randomised clinical trials in becoming a strong focus of collaboration between clinicians and epidemiologists/statisticians.

We hope that such a collaboration would lead to a stronger respect for the role played by each clinician in responding to the individual circumstances and values of their patients, and that such flexibility and humanity could be seen as a valued source of variability in practice and outcomes.

References

[1] Stark J. Future improvement in the routine monitoring of surgical performance. Bristol Royal Infirmary Inquiry, 2000 .

[2] Spiegelhalter DJ. Surgical audit: statistical lessons from Nightingale and Codman. *Statistics in Society JRSSA* 1999;162:45–58.

[3] Hunter S, Hamilton JRL, Keeton B, Anderson RH. A system for peer review for services to children with cardiac disease in the United Kingdom. Bristol Royal Infirmary Inquiry, 2000.

[4] Maternal and Child Health Consortium. Confidential Enquiry into Stillbirths and Deaths in Infancy. 6th annual report. London: Maternal and Child Health Consortium, 1999.

[5] B Botting and C Abrahams. Linking congenital anomaly and birth records. Health Statistics Quarterly 2000;8:36–40.

[6] Royal Hospitals for Sick Children, Glasgow and Edinburgh. HeartSuite 2000 System Information 2000.

[7] Berwick DM. Continuous improvement as an ideal in health care. N Engl J Med 1989;320:53–56.

[8] Schneider EC, Epstein AM. Influence of cardiac-surgery performance reports on referral practices and access to care. New England Journal of Medicine 1996;335:251–256.

[9] Spiegelhalter DJ, Evans S, Aylin P, Murray GD. Overview of statistical evidence presented to the Bristol Royal Infirmary Inquiry concerning the nature and outcomes of paediatric cardiac surgical services at Bristol relative to other specialist centres from 1984 to 1995. Bristol Royal Infirmary Inquiry: <http://www.bristol-inquiry.org.uk/brisDSAnalysis%20final.htm>, 2000 .

[10] Information for Health. Leeds: Department of Health, NHS Executive, 1998.

[11] The NHS Plan. London: The Stationery Office, 2000.

[12] An Organisation with a Memory. London: The Stationery Office, 2000.

[13] Quality and performance in the NHS: Clinical Indicators. Leeds: NHS Performance Analysis Section:
www.doh.gov.uk/nhsperformanceindicators, 2000.

[14] Clinical Outcome Indicators - 1999. Edinburgh: Clinical Resource and Audit Group, 1999.

[15] UK Cardiac Surgical Register website. <http://www.scts.org/doc/890> 2000.

[16] Central Cardiac Audit Database website.
<http://ccad3.biomed.gla.ac.uk/ccad/> 2000.

[17] Hannan EL, Racz M, Kavey RE, Quaegebeur JM, Williams R. The effect of hospital and surgeon volume on in-hospital mortality. *Pediatrics* 1998;101:963–969.

[18] de Leval MR, Francois K, Bull K, Brawn W, Spiegelhalter DJ. Analysis of a cluster of surgical failures: application to a series of neonatal arterial switch operations. *J Thoracic Cardiovascular Surgery* 1994;107:914–924.

[19] Coronary Artery Bypass Surgery in New York State, 1992-1994, Albany: New York: New York State Department of Health, 1996.

[20] Poloniecki J, Valencia O, Littlejohns P. Cumulative risk-adjusted mortality chart for detecting changes in deaths rate: observational study of heart surgery. *British Medical Journal* 1998;316:1697–1700.

[21] Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* 1997;350:1128–1130.

[22] Steiner SH, Cook RJ, Farewell VT, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000;2:–.

[23] Christiansen C, Morris C. Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* 1997;127:764–768.

[24] Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal* 1998;317:1701–1704.

[25] Burgess J, Christiansen C, Michalak S, Morris C. Medical profiling: improving standards and risk adjustments using hierarchical models. *J Health Econ* 2000;19:291–309.