Public health

Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons

Mohammed A Mohammed, K K Cheng, Andrew Rouse, Tom Marshall

During the past century, manufacturing industry has achieved great success in improving the quality of its products. An essential factor in this success has been the use of Walter A Shewhart's pioneering work in the economic control of variation, which culminated in the development of a simple yet powerful graphical method known as the control chart. This chart classifies variation as having a common cause or special cause and thus guides the user to the most appropriate action to effect improvement. Using six case studies, including the excess deaths after paediatric cardiac surgery seen in Bristol, UK, and the activities of general practitioner turned murderer Harold Shipman, we show a central role for Shewhart's approach in turning the rhetoric of clinical governance into a reality.

During the past century, manufacturing industry has achieved great success in improving the quality of its products. In industry, the definition of quality is "on target with minimum variation".[1] Reduction of variation is also a core concern in clinical governance;[2] however, there are fundamental and profound differences between the ways in which health services and industry make sense of variation. We begin with an illustration of the industrial approach to understanding and controlling variation, followed by application of this approach to health care, using six clinical governance case studies: mortality rates after paediatric cardiac surgery in Bristol, UK; mortality rates in older women treated by the general practitioner and convicted serial killer Harold Shipman; success rates of in-vitro fertilisation (IVF) treatment; neonatal deaths; prevalence of coronary heart disease in primary care; and mortality after fractured neck of femur.

Common-cause and special-cause variation

Consider a process such as writing a signature. Five of MAM's signatures are shown in the left of figure 1. Although these signatures were produced under the same conditions and by the same process, they are not identical. However, although they show variation, the variation is controlled within limits. They are all recognisably the same signature. This kind of variation suggests that a stable process produced the signatures.

In the UK National Health Service, three basic approaches are used to make sense of variation: standard setting, league tables, and hypothesis testing. Were we to compare the five signatures with a standard, some could fall below the standard. We could rank the signatures from best to worst and create a league table. A statistical test might identify one signature as being significantly different from the others. Each of these conventional approaches is inadequate because they focus our attention on the signatures that fail the test; yet, from the viewpoint of the underlying process of writing, all five signatures on the left are identical. No signature is better or worse than the others. If we want to reduce the variation between signatures, we must change the way we write all signatures, not just the ones that fail an adequate test. Thus, conventional approaches to understanding variation from a stable system can misguide us to act on individual failures rather than acting on the underlying process.

Now consider the sixth signature, on the right. It is clearly different from the others. A casual look suggests that there must be a special reason why this is so. If we want to address this kind of variation, we need to identify this special cause and prevent it from interacting with an otherwise stable process. (In this case, the signature is a forgery, attempted by TM under the same essential conditions!)

This approach categorises variation according to the action needed to reduce it. Common-cause variation is intrinsic to the process. To decrease common-cause variation, we need to act on the process. Special-cause variation is the result of factors extrinsic to the process, and its reduction therefore requires identification of and action on the special causes. The originator of these fundamental concepts was a physicist and engineer—Walter A Shewhart.[3] His pioneering work at Bell laboratories in Murray Hill, NJ, USA in the 1920s successfully brought together the disciplines of statistics, engineering, and economics, leading to the accolade: "Father of modern quality control".[4]

Shewhart devised a simple graphical method, the control chart, for discriminating between the two sources of variation, thereby guiding the user to take appropriate action. The control chart has three lines: the central line is the mean, and the upper and lower lines are termed control limits. Control limits represent the limits of common-cause variation. A data point that falls outside these control limits (or unusual patterns on the control chart) suggest a special cause. Shewhart,[3] with the aid of mathematical theory, empirical evidence, and practical concerns, advocated the use of limits set at 3 σ from the mean. The 3 σ limits are actually different from classic SD in that they measure the variability of a process over time rather than the variability of a static distribution.[5]

Figure 1: **"Lancet" signatures**

Lancet 2001; **357:** 463–67

Departments of Surgery (M A Mohammed MEng) **and Public Health & Epidemiology** (Prof K K Cheng FFPHM, A Rouse MFPHM, T Marshall MFPHM)**, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK**

Correspondence to: Dr Tom Marshall (e-mail: T.P.Marshall@bham.ac.uk)

For personal use only. Reproduce with permission from The Lancet Publishing Group.

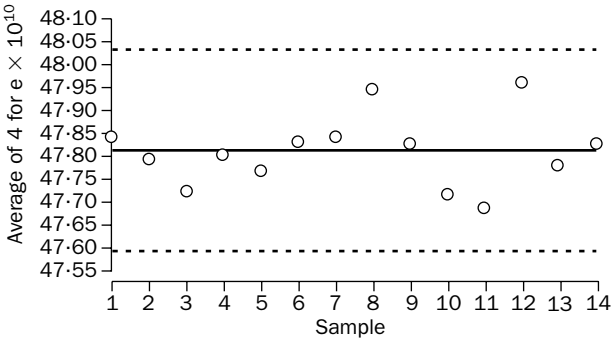

Figure 2: **Shewhart control chart of Millikan's data on charge of an electron**

There are many types of control chart. Those presented here were drawn by the method advocated by Deming[6,7] for binomial data. Such control charts can be drawn on double square-root paper (also known as probability paper) designed on the assumptions of the binomial distribution, first developed by Mostellor and Tukey.[8] The raw binomial data (x,y) are plotted on the paper, and a central line representing the mean is drawn (ie, a straight line through the origin and $\Sigma x, \Sigma y$). For more precision, a least squares line can also be computed and drawn. Since the SD on this type of paper is usefully regarded as a constant 0·5 mm, the resulting 3 σ control limits are parallel lines 1·5 mm above and below the mean.

Variation cannot be eliminated

Shewhart illustrated his concepts by applying them to the best data available to him at the time. This was a data set obtained from an experiment in which almost everything possible was done to obtain perfect results (ie, no variation)—Millikan's Nobel-Prize-winning measurements of the charge of an electron.[3] Despite Millikan's best efforts, there was substantial variation in his measurements of the charge of an electron. However, as the control chart (figure 2) of Millikan's data shows, all measurements fall within the upper and lower control limits, suggesting that his experiment was stable. To suggest to Millikan that some of his measurements were better than others, or some fell below an acceptable standard, would be absurd. Stable processes exhibit common-cause variation, which is best reduced by action on the underlying process.

Case study 1: Bristol cardiac surgery

A control chart based on data from the UK Cardiac Surgical Register of the mortality rates for children younger than 1 year old during three epochs[9] is shown in figure 3. The chart for epoch 1 will be used to explain the interpretation of a control chart.

In epoch 1, the mortality rates for nine hospitals lie within the control limits: common-cause variation. Action to reduce this variation must focus on the underlying process of care common to these nine hospitals. However, two hospitals (hospitals 11 and 7) are outside the control limits and this finding indicates that there are special causes for the variation. In hospital 11, learning why the mortality rates are high is important. To do this, we systematically look at data collection, case-mix, facilities, and quality of care. We must then take remedial action to help this hospital eliminate the special cause. In hospital 7, the mortality rate is low. It is important to find out why their results are better than other hospitals. If appropriate, we can use this knowledge to improve the results of all the hospitals.

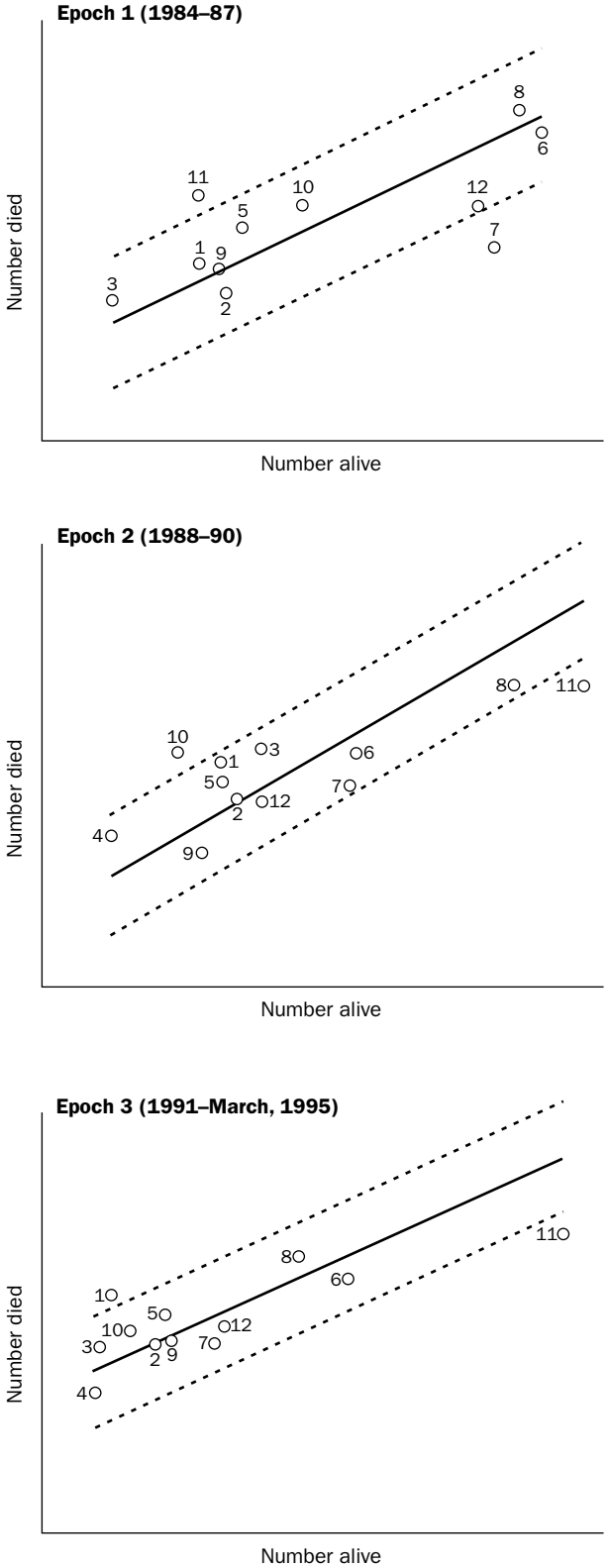

Figure 3: **Mortality for <1-year-olds after open heart surgery over three epochs**
Hospital 1 is Bristol Royal Infirmary.

In epoch 2, two hospitals (hospitals 10 and 11) show evidence of special-cause variation. Hospital 10 is in need of investigation and help to eliminate the special cause. In contrast, hospital 11 has shown remarkable improvement in its results. In epoch 1, it was above the upper control

For personal use only. Reproduce with permission from The Lancet Publishing Group.

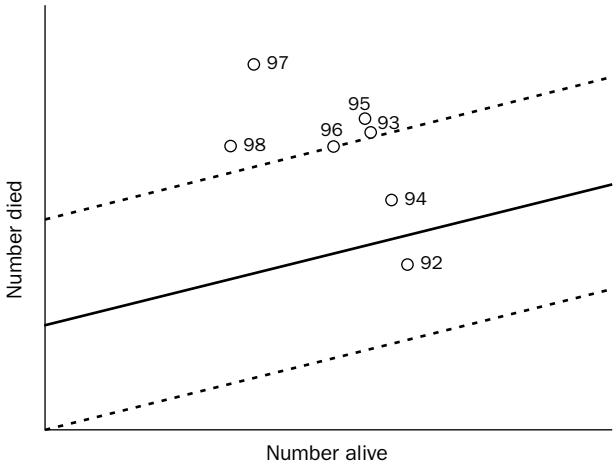

Figure 4: **Comparison of Harold Shipman's mortality for women aged 65 years or older in Thameside and Glossop during 1992–98**
The three lines indicate the background mortality for women aged 65 years or more in Thameside and Glossop (inclusive of Shipman's patients). Shipman's annual mortality rates are imposed on this.

limit, and in epoch 2 (and subsequently) it is below the lower control limit. It is also in need of investigation. Understanding why hospital 11 has made such striking progress offers an opportunity for learning, which could help the results of all. Alternatively, it may indicate that there have been important changes in the case-mix of patients treated at hospital 11.

In epoch 3, two hospitals (hospital 1 and hospital 11) show special-cause variation. Hospital 11, as in epoch 2, is below the lower control limit. Hospital 1 (Bristol Royal Infirmary) is above the upper control limit. It is in need of help to identify and eliminate the special cause for its high mortality rate.

Although external action to address concerns about paediatric cardiac surgery at Bristol Royal Infirmary took place in 1998, monitoring using the control charts could have provided a basis for action in 1987. The control charts do not only guide attention on high-mortality centres (above the upper control limit), but also clearly identify opportunities for improvement by learning from centres with low mortality rates (below the lower control limit).

Case study 2: Harold Shipman

A control chart (figure 4) of mortality rates for women aged 65 years and older in Thameside and Glossop, UK, during 1992–98[10] shows that in 1992 and 1994, Harold Shipman's mortality rates were within common-cause variation. However, during 1993 and 1995–98, his mortality rates indicated special-cause variation. To reduce special-cause variation, the special cause must be found and removed. Subsequent legal proceedings identified that special cause as being Shipman himself.

Commentators have argued that the Shipman case was not an example of poor quality of care; rather Shipman was a murderer who happened to be practising medicine.[11] This may be so, but in Shewhart's approach, murder is just one of an infinite number of special causes.

Case study 3: IVF treatment

Marshall and Spiegelhalter[12] analysed the case-mix-adjusted livebirth rate at 52 IVF clinics in the UK (n=24 739 treatment cycles, range of livebirth rate 5–24%). They concluded that league tables were

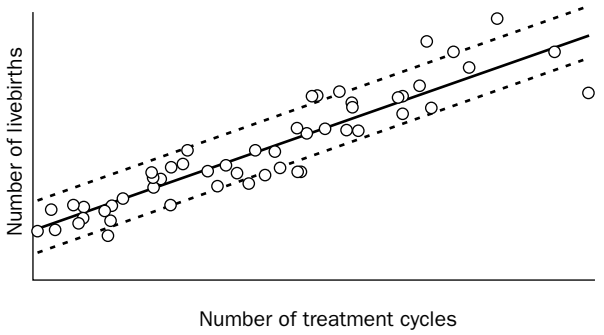

Figure 5: **Case-mix-adjusted livebirth rates from 52 IVF centres in the UK (1996)**
Clinics above and below the limits indicate special-cause variation, whereas clinics within the limits indicate common-cause variation.

unreliable. No action point emerged from their analysis. In contrast, a control chart (figure 5) with the upper and lower control limits divides the clinics into three groups with guidance for action:

- Group A—performance above the upper control limit. Find out why their results are better than other clinics. Use this knowledge to improve the performance of all the clinics.
- Group B—performance within the control limits. Make fundamental changes to the way in which IVF treatment is provided. This should be informed by lessons learned from Group A. There are no grounds for taking action in individual centres in this group.
- Group C—performance below the lower control limit. Help these centres to identify and eliminate the special causes of their poor results.

Case study 4: neonatal deaths

Parry and colleagues[13] compared mortality for nine neonatal units (n=2671 infants, mortality range 15–28%), concluding that league tables were unreliable indicators of performance. In contrast, a control chart of the neonatal data (figure 6) shows only common-cause variation, suggesting that future improvement is best sought from a fundamental change to the underlying process of care. There are no grounds for taking action on individual neonatal units.

Case study 5: prevalence of coronary heart disease in primary care

The point prevalence of coronary heart disease in a primary group consisting of 16 general practices in Birmingham, UK, was reported (private communication, Birmingham Health Authority, 1999) as 9·67% (2999/3102), with wide variation (1–38%) between practices. A control chart of these data

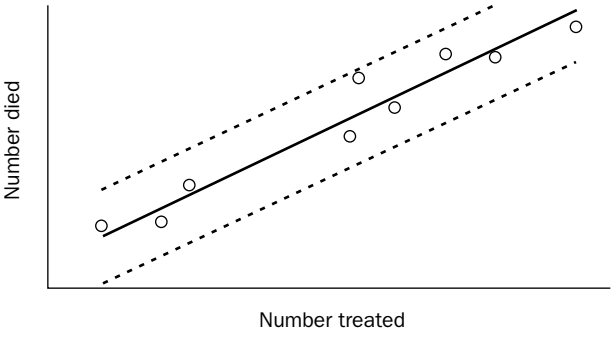

Figure 6: **Mortality in nine neonatal units**

For personal use only. Reproduce with permission from The Lancet Publishing Group.

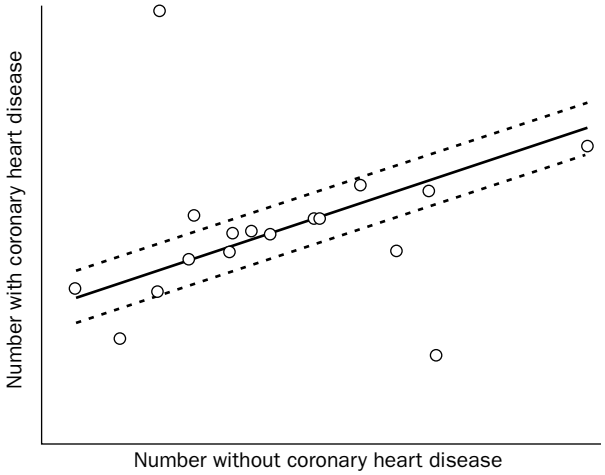

Figure 7: **Prevalence of coronary hearth disease in 16 general practices in one primary-care group in Birmingham, UK**

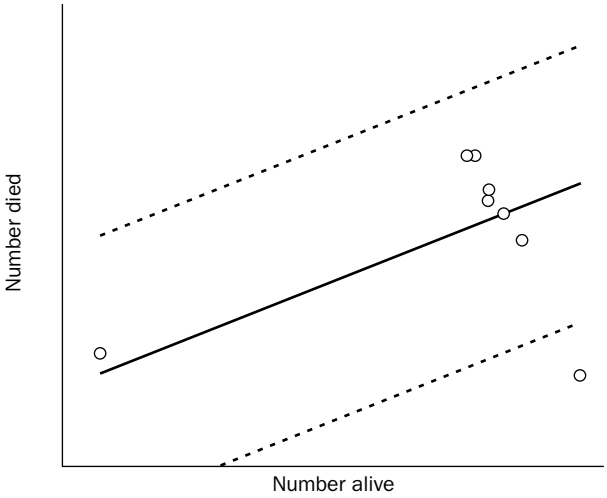

Figure 8: **Mortality after surgery for fractured hips in eight hospitals in the East Anglian Audit**

(figure 7) identifies 12 practices within control limits indicating common-cause variation. These practices should be left alone. However, five practices are outside the control limits, indicating special-cause variation. Two practices have much higher prevalence rates than expected from common-cause variation. Special-cause action needs to be taken. This should also explore the possibility of double counting. As regards the three practices below the lower control limit, they require special-cause action, which should begin with a review of the data-collection process.

Case Study 6: mortality after fractured hips

Todd and colleagues[14] compared differences in mortality after fractured hip in eight hospitals in East Anglia, UK (n=560, mortality range 5–24%). A control chart (figure 8) shows seven hospitals within common-cause variation. Improvement at these seven hospitals can only come from changing the underlying process of care for patients with fractured hip. One hospital had a very low mortality outside the limits of common cause-variation: this mortality rate is therefore likely to have a special cause. According to Todd and colleagues, this hospital employed a well-organised multidisciplinary team that sought early assessment and surgery, much of which was done by one surgeon, followed up with early postoperative mobilisation of patients. Todd and colleagues[14] were hesitant in recommending adoption of this hospital's practice, saying that "random variation" almost certainly plays a part in these findings. Shewhart's approach shows that this hospital belongs to another system beyond that attributable to random or common-cause variation. The control chart provides us with a basis for action. The model of care at this hospital should be more widely adopted perhaps after a randomised controlled trial. No action is not an option.

Discussion

These case studies illustrate an important role for Shewhart's approach to understanding and reducing variation. They demonstrate the simplicity and power of control charts at guiding their users towards appropriate action for improvement.

Actions based on Shewhart's approach are subject to two types of mistake.[14] Mistake 1 is to treat an outcome resulting from a common cause as if it were a special cause. Mistake 2 is to treat an outcome resulting from a special cause as if it were a common cause. It is impossible to reduce the frequency of both errors to zero, but what we can do is minimise the economic losses due to either kind of mistake. Shewhart argued that variation from stable processes lies within limits which—combining mathematical theory, empirical evidence, and pragmatism—can be most usefully set at 3 σ limits from the mean.

Some may regard limits of 3 σ as too wide a range for health care. The use of a narrower range, say 2 σ, might seem more appealing. But there is a need for caution. First, as shown by the electron data, stable systems can and do produce data beyond 2 σ limits. So we will be guided to look for trouble more often then it actually exists (mistake 1). Given the culture of blame in health services, we risk making matters worse, especially when the person closest to the failure is held to be responsible. Furthermore, the case studies used here show that the 3 σ limits are adequate to find special-cause variation in practice.

Perhaps it would be optimistic to suggest that use of control charts could prevent the recurrence of tragic and unfortunate episodes such as Bristol or Shipman. What is clear is that analysing data with an understanding of common-cause and special-cause variation provides health services with a basis to act. There is an axiom that the purpose of data is action.[15] Each of the above case studies is based on data available at the time of the events. In each case, little or no action was taken at the time. Why? We believe this is largely because the current methods for understanding variation in health services provide little or no guidance for action. One prominent advocate of Shewhart's method was so convinced of this that he wrote: "Tests of significance, t-test, chi-square, are useless as inference—i.e., useless for aid in prediction. Test of hypothesis has been for half a century a bristling obstruction to understanding statistical inference".[15] At least we should seriously question the role of conventional statistical analysis in clinical governance.

Shewhart's concepts provide a sober antidote to the plague of league tables. Under stable conditions, league tables are unreliable and their guidance is equally unreliable.[16] Action based on their guidance is likely to be misguided, resulting in tampering and making matters worse.[15]

The suggestion that Shewhart's work might be useful in health care is not new.[17] The technique is in current

For personal use only. Reproduce with permission from The Lancet Publishing Group.

use in the USA. There it has found applications in improving the quality of ambulatory clinical care;[18] the analysis of longitudinal variations in trauma mortality;[19] reducing hospital-acquired infection;[20] and identifying changes in the historical pattern of disease which require public-health investigation.[21] Other charting techniques, notably CUSUM (cumulative sums), have been used for longitudinal analysis of surgical mortality.[22–24] These techniques can also detect changes in performance. However, CUSUM charts require the setting of a target, which is not always possible in clinical medicine, and the technique has not routinely been employed for comparative analysis of variation across health-care providers. Control charts are generally straightforward to produce and easy to interpret.

The era of clinical governance offers immense opportunities. In the past, those in possession of data might have opted for inaction or called for better data. Recent high-profile cases have contributed to conditions where the tendency for action will be more frequent. The case for the control chart to guide action has been presented. Its guidance has proved immensely useful to industry over the past 50 years;[1,6,15] it is time for it to be integrated into clinical governance.

Contributors

Mohammed A Mohammed originated the idea, provided case study examples and analyses, and contributed to the writing of the paper; K K Cheng conceived the idea of the paper and contributed to the writing of the paper; Andrew Rouse provided a case study example and analysis, and contributed to the writing of the paper; and Tom Marshall contributed to the design and writing of the paper.

Acknowledgments

We thank Henry Neave for introducing MAM to the Deming management method; David Kerridge for providing expert help and guidance throughout this work, particularly on the theory and practice of control charts; and Malcolm Gall for contributing the initial control chart of Harold Shipman's mortality data from *The Times*.

References

1 Wheeler D. Advanced topics in statistical process control. The power of Shewhart's charts. Knoxville: SPC Press Inc, 1995.
2 Scally G, Donaldson LJ. The NHS's 50 anniversary. Clinical governance and the drive for quality improvement in the new NHS in England. *BMJ* 1998; **317:** 61–65.
3 Shewhart WA. Economic control of quality of manufactured product. New York: D Van Nostrand Company, 1931. (Reprinted by ASQC Quality Press, 1980).
4 American Society for Quality www.asq.org/about/history/shewhart.html, accessed Jan 2001.
5 Shahian DM, Williamson WA, Svensson LG, Restuccia JD, D'Agostino RS. Applications of statistical quality control to cardiac surgery. *Ann Thorac Surg* 1996; **62:** 1351–59.
6 Deming WE. Out of the crisis. Massachusetts: Massachusetts Institute of Technology, 1986.
7 Deming WE. Sample design in business research. New York: Wiley Classics Library Edition, 1990.
8 Mostellor F, Tukey J. The uses and usefulness of probability paper. *J Am Stat Assoc* 1949; **44:** 174–212.
9 Spiegelhalter DJ, Evans S, Aylin P, Murray G. 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. www.bristol-inquiry.org.uk/Documents/statistical%20overview%20report.pdf www.bristol-inquiry.org.uk/Documents/fig61.pdf, accessed Jan 2001.
10 Duce R, Jenkins R. GP probably killed more than 100. *The Times*, Feb 1, 2000.
11 Anonymous. Shipman was a criminal and not a poorly performing doctor. *BMJ* 2000; **320:** 1734.
12 Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ* 1998; **316:** 1701–05.
13 Parry GJ, Gould CR, McCabe CJ, Tarnow-Mordi WO. Annual league tables of mortality in neonatal intensive care units: longitudinal study. *BMJ* 1998; **316:** 1931–35.
14 Todd CJ, Palmer C, Camilleri-Ferrante C, et al. Differences in mortality after fracture of hip: the East Anglian audit. *BMJ* 1995; **310:** 904–08.
15 Deming WE. The new economics, second edn. Massachusetts: Massachusetts Institute of Technology, 1994.
16 Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J R Stat Soc Series A* 1998; 159; **3:** 385.
17 Berwick DM. Controlling variation in health care: a consultation from Walter Shewhart. *Med Care* 1991; **29:** 1212–25.
18 Allison JJ, Calhoun JW, Wall TC, et al. Optimal reporting of health care process measures: inferential statistics as help or hindrance? *Managed Care Q* 2000; **8:** 1–10.
19 Clark DE, Cushing BM, Bredenberg CE. Monitoring hospital trauma mortality using statistical process control methods. *J Am Coll Surg* 1998; **186:** 630–35.
20 Benneyan JC. Statistical quality control methods in infection control and hospital epidemiology, part I: introduction and basic theory. *Infect Control Hosp Epidemiol* 1998; **19:** 194–214.
21 Williamson GD, Weatherby Hudson G. A monitoring system for detecting aberrations in public health surveillance reports. *Stat Med* 1999; **18:** 3283–98.
22 de Leval MR, Francois K, Bull C, Brawn W, Spiegelhalter D. Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg* 1994; **107:** 914–23.
23 Poloniecki J, Valencia O, Littlejohns P. Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *BMJ* 1998; **316:** 1697–700.
24 Steiner SH, Cook RJ, Farewell VT. Monitoring paired binary surgical outcomes using cumulative sum charts. *Stat Med* 1999; **18:** 69–86.

For personal use only. Reproduce with permission from The Lancet Publishing Group.