APPENDIX E

‘Following Shipman: a pilot system for monitoring mortality rates in primary care’

Reproduced with Permission from The Lancet

Public health

Following Shipman: a pilot system for monitoring mortality rates in primary care

Paul Aylin, Nicky Best, Alex Bottle, Clare Marshall

As part of the investigations into the crimes of Harold Shipman, it has become clear that there is little monitoring of deaths in general practice. By use of data on annual deaths at family physician and practice level for five English health authorities for 1993–99, we investigate whether cumulative sum charts (a type of statistical process control chart) could be used to create a workable monitoring system. On such charts, thresholds for deaths can be set, which, if crossed, may indicate a potential problem. We chose thresholds based on empirical calculations of the probabilities of false and successful detection after allowing for multiple testing over physicians or practices. We also statistically adjusted the charts for extra-Poisson varition due to unmeasured case mix. Of 1009 family physicians, 33 (including Shipman) crossed the alarm threshold designed to detect a 2 SD increase in standardised mortality, with 97% successful detection and a 5% false-alarm rate. Poor data quality, plus factors such as the proportion of patients treated by these physicians in nursing homes or hospices are likely explanations for most of these additional alarms. If used appropriately, such charts represent a useful tool for monitoring deaths in primary care. However, improvement in data quality is essential.

As part of the investigations into the crimes of Harold Shipman, a UK family physician who has been imprisoned for the murder of many of his patients, it has become clear that there is little formal monitoring of deaths in general practice.[1] However, if an acceptable and workable method of monitoring mortality rates can be devised, implementation of such a system might be beneficial. Baker and colleagues[2] have suggested that monitoring mortality in general practice should be able to detect illegal behaviour, but also could help to inform the quality of clinical care and maintain public trust.

We have assessed the feasibility of setting up a system for the routine surveillance of mortality data at individual family physician and practice levels. We focused on the data requirements and statistical issues involved, especially between-unit variation due to the net effect of many small unmeasured factors (eg, case mix, data errors), and the difficulty of multiple testing over units and over time. We use data from a retrospective pilot exercise, commissioned by the Shipman Inquiry,[3] to link national death registration records to lists of primary-care patients held in five health authority systems. Statistical process control (SPC) charts, which have been advocated for use in clinical governance,[4–6] are discussed and applied to the data to illustrate the practical application of such techniques in this context.

Statistical Issues

A monitoring system must be designed to quickly detect unusual variations in the underlying mortality rate in any unit. To design such a system requires that expected or acceptable mortality rates be defined, corresponding to what is termed the in-control process in control-chart studies. Criteria must also be decided to define when the observed mortality is sufficiently different from that expected to warrant special attention, corresponding to the mortality rates being out of control. Units that meet the latter criteria should be followed up to seek explanations for the causes of the variability.

Various statistical methods can be used to detect unusual outcomes or changes in the level of an underlying process, and have been applied in the context of surveillance and performance monitoring.[7–9] However, many methods are designed as one-off tests to be done only after all the data have been collected.[7,9] The distinctive feature of a prospective monitoring system is that data accumulate over time and the analysis is repeated at every time point. This approach is known as sequential analysis and is widely used in industrial quality control[10] and for determing stopping rules for clinical trials.[11]

SPC charts are among the most widely used methods for sequential analysis. Various types of SPC chart have

Key features of SPC charts

Test statistic calculated for the unit at each time point
This statistic is typically a function of the difference between the observed outcome at a given time and that expected under the in-control distribution. The statistic may also depend on previous values of these residuals, leading to a cumulative sum.

Predefined alarm threshold
If the test statistic exceeds the threshold at some time *t*, a warning or alarm is triggered and the chart is said to signal that the process being monitored has become out of control.

Some measure of the performance of the chart
The chart's ability to detect when the underlying process is truly in and out of control must be measured. Such chart performance measures take the place of the more familiar type I (false positive) and type II (false negative) error rates in ordinary significance tests and are used, in all but the Shewhart chart, to inform the choice of alarm threshold *h*.

Lancet 2003; **362:** 485–91. Published online July 29, 2003
http://image.thelancet.com/extras/03art6478web.pdf
See Commentary page 417

Department of Epidemiology and Public Health, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PF, UK
(P Aylin FFPHM, N Best PhD, A Bottle PhD, C Marshall PhD)

Correspondence to: Dr Paul Aylin
(e-mail: p.aylin@imperial.ac.uk)

been developed. All charts share the key features listed in the panel (further details of the most commonly used SPC charts are given at http://image.thelancet.com/extras/03art6478webappendix.pdf), and the statistical features of these charts are discussed in depth by Marshall and colleagues[12] and Sonesson and Bock.[13] Such methods have been used in public-health surveillance[14,15] and by individual hospitals, such as for monitoring of surgical mortality rates[16] and hospital-acquired infections.[17] Their use has also been suggested for clinical governance to show when the performance of an individual or institution may have crossed a prespecified warning or alarm threshold;[4] had they been in place at the time, they might have shown the unacceptably high paediatric cardiac surgical mortality rates at Bristol Royal Infirmary and among Harold Shipman's patients.[6,18] However, the origin of SPC charts in industrial quality control must be borne in mind. Although the characteristics of an industrial process are clearly vastly different from those of a health outcome process, such as surgical or general practice mortality rates, the implications of these differences for the use of SPC charts in clinical governance are not widely appreciated.

Most industrial processes are well characterised in the sense that, when a process is in control, the only source of variation is random. By contrast, health outcome processes are far more complex. Even when in control, the process is subject to many non-random sources of variation, particularly changes in case mix, that can dominate chance fluctuations. An efficient SPC chart for clinical governance must, therefore, specify an acceptable in-control performance level and an acceptable amount of variation about this level. Partial adjustments of the in-control outcome rate for measured case-mix variables can frequently be made.[16,19] However, lack of data, an appropriate risk-adjustment scheme, or both, means that the in-control variance should also reflect the inevitable within-unit (individual or institution) and between-unit variation in outcomes due to unmeasured differences in case mix and other factors beyond the control of the unit. Hence the in-control variance should be larger than that assumed to be due to chance alone—commonly termed overdispersion. Failure to allow for overdispersion of the in-control process will result in an unnecessarily high false-alarm rate. However, specification of an appropriate amount of overdispersion can be difficult. Ideally, the amount should be estimated from historical data on units known to have had acceptable performance levels. Alternatively, it could be chosen subjectively by taking into account the degree of variation in outcomes that would be expected for units with extreme combinations of case mix.

A second challenge arises in the context of a national primary-care surveillance system if a central audit body is to examine multiple SPC charts from several units. The alarm threshold for an SPC chart is chosen based on a trade-off between the expected true and false alarm rates for different thresholds. Although the standard measures used to quantify error rates for SPC charts acknowledge the multiple testing of data over time for one unit,[20] multiple testing over many units must be taken into account when assessing the chart error rates to set appropriate alarm thresholds.

The traditional approach to multiple significance testing controls the overall false-positive (type I) error rate. For example, the widely used Bonferroni correction[21] ensures that if M tests are done and all M null hypotheses are true, the probability of falsely rejecting at least one null hypothesis is less than or equal to the specified overall type I error rate. Although Spiegelhalter and colleagues[18] have tentatively recommended using Bonferroni corrections in clinical monitoring, we believe that this approach is not wholly appropriate. One problem is that the type I error rate for a sequential analysis is not constant, but increases with the length of the surveillance period (the probability of eventually signalling an alarm is 1 for all sequential tests). More fundamentally, however, we argue that we are not so much concerned with controlling the probability of getting at least one false alarm out of the M units being monitored, as with estimating the proportion of all alarms detected that are false—the latter has been termed the false detection rate (FDR).[22] By the same argument, the successful detection rate (SDR) could be estimated, that is, the proportion of true out-of-control units successfully identified—a concept somewhat analogous to the power of the surveillance system. In the context of multiple SPC charts, the FDR and SDR may be estimated with computer simulation methods; we have reported values of FDR and SDR at time t—ie, the expected proportion of false alarms among all alarms occurring by time point t after the start of monitoring, and the proportion of true out-of-control units successfully identified by time t, respectively—for various choices of alarm threshold for the cumulative sum control chart.[12,23]

An important feature of the FDR is that it depends on the true (but unknown) proportion of out-of-control units—the smaller the proportion of out-of-control units, the higher the expected proportion of alarms that are false. We have investigated methods for estimating the true number of out-of-control units,[12] although this topic requires further investigation. However, in the context of mortality rates in primary care, we would not expect more than 5–10% of all family physicians or practices to have truly unacceptable mortality rates. Therefore, for illustration we report the FDR, assuming that either 5% or 10% of units are out of control. The expected FDR increases with the length of surveillance and, like the type I error for a single chart, will eventually equal 1. The expected SDR does not depend on the proportion of true out-of-control units, but does increase with length of time that has elapsed since the specific processes shifted from in control to out of control.[12,23]

As already noted, the FDR and SDR can be used to help choose an appropriate alarm threshold for the SPC chart. We make no specific recommendation for this choice. However, within the context of monitoring mortality rates in primary care, a scenario could be anticipated whereby the central audit body would be willing to tolerate several false alarms to ensure a sufficiently high SDR. This trade-off can be formalised within a statistical decision-making framework.[24] For example, the relative cost of failing to detect a family physician whose patients' mortality rate is out of control may be deemed twice as great as falsely detecting a family physician whose patients' mortality rate is truly in control. The alarm threshold would be chosen to keep the overall loss to a minimum, where the latter is expressed as an appropriate function of the above two-to-one cost ratio and the FDR and SDR for various thresholds and time points. Such cost-benefit calculations could also take into account the practical and financial costs of implementing a system to follow-up any alarms.

Data linkage

The current death registration process does not record the deceased's family physician. Only the name of the certifying doctor is recorded. To obtain deaths by practice

or family physicians, we extracted 7 years of mortality data (1993–99) from the statutory death register held by the Office for National Statistics. We linked the data with general practices' lists of patients held on five English health authority information systems (including that in which Shipman's practice was situated, West Pennine). With use of the National Health Authority Information System (NHAIS), deaths were linked through patients' NHS numbers or, if not present, date of birth, sex, and postcode, to provide a family physician, practice, and senior partner code on each mortality record, together with the NHAIS date of death field (for quality comparisons). Once linked, personal identifiers were removed from the records. The National Health Service Information Authority also provided list sizes from quarterly capitation figures, broken down by three broad age-groups (0–64, 65–74, and ≥75 years) for each family physician and practice by year.

Numbers of deaths and populations were aggregated by the family physician's General Practitioner National Code (the national identifier used by the NHS) and the national practice code. A few deaths could be allocated only to a family physician and not to a practice. These deaths were excluded from analyses of practices. Family physicians whose practices were near the boundary of a health authority had only part of their list recorded by that health authority's system if some patients were resident in neighbouring health authorities. We were unable to find out whether lists of patients were small because of such positioning or simply because few patients were registered. For the purpose of showing the usefulness of these data for monitoring, we based all analyses on a subset of practices and family physicians from among those studied that we thought had almost complete data (we used a cut off of >1000 recorded population): according to the National Database for Primary Care Groups and Trusts, in 2001 the average family physician's list size was just less than 2000 patients, and fewer than 4% of family physicians had a list size less than 1000. We therefore selected units on the criteria that they served one of the five pilot health authorities and had a recorded population of at least 1000 within that health authority for every year of the study period.

We used indirect standardisation to estimate the expected (in-control) mortality count for each year for each family physician and practice. Unfortunately, adjustment for case mix was limited to age (0–64, 65–74, and ≥75 years) because of the lack of information available in the denominator (quarterly capitation) data. Local reference rates were derived from the age-specific and year-specific mortality rates for the relevant health authority, and the list size for each unit (family physician or practice) was taken as the average quarterly capitation per year, since not all units had data on lists of patients available for all four quarters in a year.

Construction of control charts

A common statistical assumption is that mortality counts follow a Poisson distribution with mean given by the indirectly standardised expected count. However, since we were able to make only limited adjustment of the expected (in-control) mortality counts for case mix, there are probably many unmeasured risk factors leading to variation in the observed data other than that due to random Poisson variation. We obtained estimates of the amount of extra-Poisson variation (overdispersion) in the mortality counts at family physician and practice level each year, using quasi-likelihood estimation methods[25] applied to the data for the five health authorities in the pilot study.[23] In an attempt to partly overcome overestimation of the level of extra-Poisson variability due to the presence of out-of-control units, we use the median value of overdispersion across the 7-year monitoring period in all subsequent analysis. As already noted, however, basing these overdispersion estimates on a separate historical dataset containing only family physicians or practices known to be in-control would have been preferable. Because of the nature of the pilot study, however, no such data were available. It should be possible to identify relevant historical data in any future application of these methods to prospective monitoring of mortality data.

The annual mortality counts for each unit (family physician or practice) were transformed to roughly follow a standard normal distribution by application of the Poisson half-sum transformation,[26] and then dividing by the square root of the overdispersion factor to further standardise the variance of the transformed counts. This approach yielded a standardised residual for each unit and year that is roughly normally distributed, with mean equal to zero and SD equal to one when the unit's mortality rate is in control. This residual is referred to subsequently as the standardised excess mortality. Graphic checks, such as qq plots, suggested that the normality assumption for the standardised excess mortalities was reasonable.[23]

The standardised excess mortalities were used to construct normal log-likelihood ratio cumulative sum charts for each unit.[23] A feature of these charts is that we must specify in advance how large an increase in mortality rate we are interested in detecting; this limit defines what we deem to be an out-of-control process that warrants special attention. Here, we arbitrarily chose to construct cumulative sum charts to detect an out-of-control mortality rate process with mean standardised excess mortality equal to, variously, K=1, 2, or 4. This value corresponds to the detection of units whose mean number of deaths is K SD above that expected for an in-control unit, where the SD reflects the acceptable within-unit and between-unit variability in mortality due to chance fluctuations and the effect of unmeasured case-mix. Cumulative sum charts can also be constructed to detect a decrease in mortality rates (ie, to identify units with exceptionally good performance) although we have not covered such charts in this report.

Application of charts

The Office for National Statistics provided the NHSIA with 281 777 mortality records from the five pilot areas for the years 1993–99. The success of linking mortality data to registers of lists of patients varied over time and between health authorities. Overall, 92% of mortality records were linked with data on list of patients. The proportion of mortality records successfully linked in all but the West Pennine health authority was similar, with the match rate for the other four health authorities ranging from 90% to 97% for the 1993 data and improving to between 96% and 99% for the 1999 data. Matching was less successful (<60%) for the West Pennine health authority before 1999. That health authority was created by the merger of two family health service authorities, Oldham and Tameside. In database terms, the Oldham live-patient data were moved over to the Tameside database, but the data for patients who had died were removed from the list. This merger took place from late 1998 to early 1999. As a result, the West Pennine database contains data for all West Pennine patients after early 1999 but only for ex-Tameside patients before this date. For all five health authorities, there was a noticeable

jump in match rates when NHS numbers started to appear in the Office for National Statistics mortality data. We identified 854 practice and 2705 family physician codes. We excluded units that did not serve one of the five health authorities under study, leaving 553 practices and 1899 family physicians. We also excluded units that did not satisfy our criterion of at least 1000 patients on their list in every year. In 1999, 524 practices and 1583 family physicians had lists greater than 1000 patients, but only 101 (12%) practices and 1009 (37%) family physicians had such lists for all 7 years.

Mortality varied widely between units at both family physician and practice levels, despite some adjustment for patients' age. For practices, the median annual number of deaths was 36 (range 0–212); standardised mortality ratios varied from 0 to 2·26 (SD 0·45). For family physicians, the median annual number of deaths was 21 (0–78), and standardised mortality ratios varied from 0 to 2·45 (0·32). Median estimates of overdispersion suggested that, after adjustment for age, the variability in mortality rates between practices was almost 3·5 times greater, and between physicians 2·0 times greater, than that expected by chance alone.

To help place these estimates of variability in context, we considered how Harold Shipman's mortality rates fit into this pattern of variation. During the early 1990s, Shipman recorded around ten more deaths per year than expected, rising to nearly 40 excess deaths in the year before his arrest. Table 1 shows the number of SD of the in-control mortality process, after adjustment for overdispersion, that correspond to ten or 40 excess deaths for a unit at family physician and practice levels. These numbers give the maximum value of K that would have to be used to construct a cumulative sum chart that could detect an out-of-control unit with an excess number of deaths per year similar to Shipman's. For example, if monitoring annual mortality at family physician level with the aim of detecting units with ten (40) excess deaths, we would need to construct cumulative sum charts for each family physician to detect an out-of-control increase in standardised excess mortality of K=1·7 (K=6·6).

Cumulative sum charts were run on the data at both family physician and practice level. The numbers of units signalling an alarm during the monitoring period, together with the FDR and SDR for the corresponding charts, are shown in tables 2 and 3. As an example of how to interpret these tables, we focus on the family physician level results in table 2 and look at the detection of an increase in mortality of 2 SD higher than the in-control mean (ie, K=2·0). The fifth row of table 2 shows that when the alarm threshold was set at h=3, 33 (3·3%) family physicians signalled at some time within the 7-year monitoring period. The estimated 7-year FDR (FDR_7) for this chart shows that between 5·2% and 2·5% of these 33 signals are expected to be false alarms if between 5·0% and 10·0% of the 1009 family physicians have mortality rates that are truly out of control. If fewer than 5·0% of family physicians are truly out of control, the proportion of false alarms is expected to be higher than 5·2%. By use of this particular chart, we would also expect to correctly detect more than 96% of the family physicians whose mortality rates had been truly out of control for the 7 years of monitoring (SDR_7). Increasing the size of the out-of-control excess mortality to be detected to K=4·0 more than halves the number of family physicians who signal at a given alarm threshold, since we are now aiming to identify much more extreme deviations in performance. The estimated detection rates for these charts show very few family physician alarms are expected to be false, and that they should include nearly all family physicians whose mortality rates have been truly out of control since the start of monitoring.

We also specifically examined West Pennine NHAIS data to find out whether Harold Shipman could have been identified by use of the methods discussed in this report. Shipman's General Practitioner National Code number was revealed to us only after the main analysis. By use of a cumulative sum chart with K=2·0 (thus aiming to detect family physicians with mortality rates for patients that are at least 2 SD higher than the acceptable level) Shipman is

	Family physician	Practice
Mean number of deaths observed	20	49
Overdispersion on factor	1·84*	3·42*
Number of deaths corresponding to 1 SD	6·1	12·9
Number of SD corresponding to ten extra deaths	1·7	0·8
Numbers of SD corresponding to 40 extra deaths	6·6	3·1

*Assumes deaths follow overdispersed Poisson distribution, with this level of overdispersion.

Table 1: **Relation between actual number of excess deaths (observed *vs* expected) and SD of in-control distribution of standardised residual mortality for units at each level of aggregation**

	Number of units whose cumulative sum chart crossed threshold (n [%])	Multiple-chart performance measures		
		SDR_7 (%)	FDR_7 (%)	
			5% of units truly out of control*	10% of units truly out of control†
Shift K/threshold h				
1/2	138 (13·7)	88·9	75·6	59·5
1/3	83 (8·3)	74·7	48·9	31·2
1/5	37 (3·7)	41·2	8·8	4·4
2/2	52 (5·1)	99·1	31·2	17·6
2/3	33 (3·3)	96·6	5·2	2·5
2/5	23 (2·3)	82·4	0·2	0·1
4/2	16 (1·6)	>99·9	<0·01	<0·01
4/3	12 (1·2)	>99·9	<0·01	<0·01
4/5	8 (0·8)	>99·9	..	..

SDR_7=Proportion of out-of-control units successfully detected within 7 years. FDR_7=Proportion of signals observed in 7 years that are false alarms. *Assuming 5% of 1009 units truly out of control. †Assuming 10% of 1009 units truly out of control

Table 2: **Proportion of 1009 family physicians signalling alarm at any time between 1993 and 1999 for different values of K (increase in standardised excess mortality) and h (alarm threshold)**

	Number of units whose cumulative sum chart crossed threshold (n [%])	Multiple-chart performance measures		
		SDR_7 (%)	FDR_7 (%)	
			5% of units truly out of control*	10% of units truly out of control†
Shift K/threshold h				
1/2	14 (13·9)	88·9	75·6	59.5
1/3	10 (9·9)	74·7	48·9	31·2
1/5	7 (6·9)	41·2	8·8	4·4
2/2	6 (5·9)	99·1	31·2	17·6
2/3	5 (5·0)	96·6	5·2	2·5
2/5	3 (3·0)	82·4	0·2	0·1
4/2	0	>99·9	<0·01	<0·01
4/3	0	>99·9	<0·01	<0·01
4/5	0	>99·9	<0·01	<0·01

SDR_7=Proportion of out-of-control units successfully detected within 7 years. FDR_7=Proportion of signals observed in 7 years that are false alarms. *Assuming 5% of 101 units are truly out of control. †Assuming 10% of 101 units are truly out of control.

Table 3: **Proportion of 101 practices signalling at any time between 1993 and 1999 for different values of K (increase in standardised excess mortality to be detected) and h (alarm threshold)**

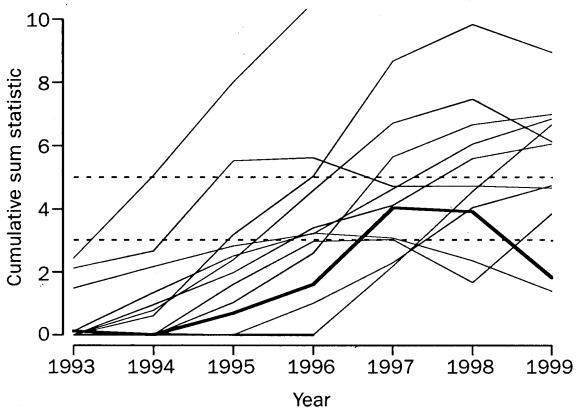

Figure 1: **Cumulative sum charts for 12 family physicians signalling at any time between 1993 and 1999**
Charts designed to detect 4 SD increase in standardised excess mortality (K=4·0) with an alarm threshold of h=3 (h=5 is also shown). Bold line=Harold Shipman's cumulative sum chart.

one of the 33 family physicians that signal at the h=3 alarm threshold. He is also one of the 23 that signal at an alarm threshold of h=5. If instead we use a chart designed to detect a K=4·0 SD increase in standardised excess mortality, then 12 family physicians signal at an alarm at threshold h=3, one of whom is Harold Shipman. For this chart, we expect to detect more than 99·9% of truly out-of-control family physicians within 7 years of monitoring, with very little chance (<0·01%) of sustaining any false alarms. Figure 1 shows the cumulative sum charts for these 12 family physicians. Harold Shipman's chart crosses the h=3 threshold for the first time in 1997. However, Shipman's chart does not signal at the h=5 threshold, although charts for eight other family physicians do, which suggests that Shipman's cumulative sum chart is by no means the most extreme. Although Shipman ceased practising after his arrest in 1998, his cumulative sum chart continues beyond this time because the locum who took over his practice used his General Practitioner National Code number. This anomaly raises the question of what to do to a chart after a signal. One option would be to reset the chart to zero, or perhaps to a head-start value greater than zero, which would put the unit on a kind of probation, and continue monitoring.

Figure 2 shows Shipman's chart again, along with those of two other selected family physicians. Also given are plots of the annual standardised excess mortalities (assumed to follow a standard normal distribution when in control) for these family physicians over time, showing the data on which the cumulative sum charts are based. This figure highlights several important features of the cumulative sum procedure. If we focus on Harold Shipman, in 1994 the observed number of deaths among his patients were fewer than those expected. His cumulative sum chart later rises as the observed number of deaths begins consistently to exceed that expected, crossing the h=3 alarm threshold in 1997, coincidently the same year in which his annual standardised excess mortality first exceeds 4 SD higher than the in-control mean. For the first other family physician, the standardised excess mortality increases sharply in 1996 but the cumulative sum chart rises only slightly, dampened by the lack of cumulative evidence to suggest an important shift in the process. The standardised excess mortality for this family physician drops back down in the following year, which suggests that the previous year's rise

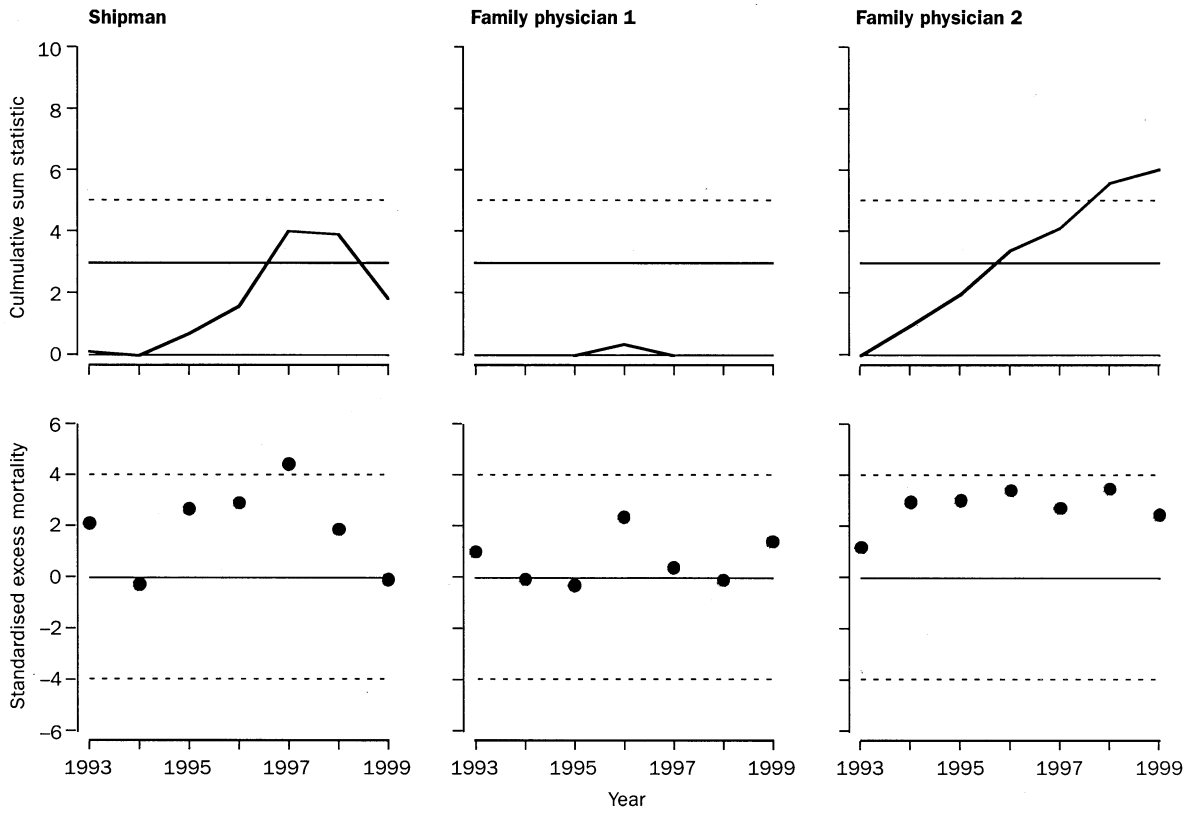

Figure 2: **Cumulative sum charts and traces of standardised excess mortality for Harold Shipman and two other family physcians**
Charts designed to detect 4 SD increase in standardised excess mortality (K=4·0). Thresholds of h=3 and h=5 shown.

may be random variation. For the second other family physician, the standardised excess mortality never exceeds 4 SD higher than the mean in any year. The cumulative chart, however, crosses both alarm thresholds (h=3 and h=5) as more and more evidence is accumulated over time to suggest that this family physician's mortality rates do not fall within the range of acceptable variation for an in-control process.

How charts compare

Various methods for health-care surveillance have been proposed,[6,27,28] but none seems to have dealt effectively with the monitoring of multiple units over multiple points in time. Nor has the difficulty of how to adjust SPC charts for the inevitable overdispersion in routinely collected health outcome data been previously considered. We have shown the usefulness of log-likelihood ratio cumulative sums for monitoring mortality in primary care with use of real data for a large number of units over 7 years. We were able to adjust statistically for multiple sources of acceptable variation in the mortality data, in particular for unmeasured case-mix factors. We have shown also that alarm thresholds can be calibrated to vary the FDRs and SDRs of a chart, allowing for multiple comparisons across units as well as over time. Our results indicate that log-likelihood ratio cumulative sum charts could be an effective tool for monitoring annual mortality at family physician and practice level, and would have been capable of detecting Shipman (along with a small number of physicians or practices) if they had been in place at the time.

Cumulative sum charts are not the only SPC charts that could have been used. Shipman could have been detected by use of alternatives such as the Shewhart chart[6] or the sequential probability ratio test,[18] although these methods have not been studied in the context of routine monitoring of many hundreds of family physicians or practices. However, irrespective of which SPC chart is used, the issues of overdispersion and multiple testing across units need to be addressed.

A comparative review of the various types of SPC chart for monitoring health outcome processes is given by Sonnesson and Bock,[13] and we summarise the main advantages and disadvantages of these methods in the web appendix (http://image.thelancet.com/extras/03art6478webappendix.pdf). We argue that Shewhart charts are not the most suitable method for monitoring mortality if the goal is to detect evidence of systematic poor-quality care, as well as of illegal behaviour. The Shewhart chart test statistic is calculated independently at each monitoring time, with use of only the most recent observation (or mean of the observations since the last monitoring time), and so takes no account of accumulating evidence of sustained poor performance. The chart test statistics for most other types of SPC chart, including cumulative sum and sequential probability ratio test, are based on a cumulative sum of outcomes over current and previous monitoring time points. These sums are better able to detect units with small to moderate sustained excess mortality over time, as well as units with sudden large increases in mortality, than non-cumulative charts. The choice between the various types of cumulative sum chart is subtle, and requires further empirical investigation before a definitive recommendation can be made. However, we would argue that cumulative sums are preferable to sequential probability ratio test charts since the cumulative sum chart statistic is bounded below by zero, whereas the sequential probability ratio test chart statistic becomes increasingly negative if the unit remains in control. This effect allows the unit to build up credit for good past performance, but has the disadvantage that any subsequent worsening in performance can be masked, at least during the early period after a change.

Interpretation of charts

Patients' mortality is clearly highly variable at family physician and practice levels. This variability is much greater than would be expected by chance alone. Key explanations for this finding include inadequate adjustment for case-mix and poor data quality yielded from the pilot exercise to link primary care populations and mortality at family physician level. The limited success of the latter was largely due to missing NHS numbers on the Office for National Statistics mortality records for earlier years, poor denominator data in one health authority, and some missing data as a result of health authority mergers. However, linkage improved over time, and we would expect any future surveillance exercise using national data to meet with more success. Now operational, the NHS Strategic Tracing Service[29] is supplied with downloads from primary-care Trust information systems. This service will assist NHS organisations trying to trace patients and may be useful in gaining national data on death.

Any exercise in linking the current mortality record with data on lists of patients can only result in rates for patients registered with a particular family physician or practice and cannot provide information on the identification of the certifier, since it is not currently coded in national mortality data. An alternative to data linkage is to record the identity of the family physician and the certifying doctor at registration with a unique identity code. Whatever new information is collected on the death certificate, denominators would still be required and these may be difficult to ascertain if the analysis is by certifier rather than registered family physician.

We were unable to adjust for case-mix factors other than age because of limited historical denominator data. The attribution dataset, although not available to us for historic analysis, could be used by any surveillance system in the future. This dataset is compiled from NHAIS and contains data by ward (thus providing a link to census data for deriving measures of socioeconomic deprivation), family physician, sex, and a finer age breakdown. Other case-mix factors that might be taken account of include the number of nursing homes or hospices in which patients are treated by a practice or specialisation of an individual family physician, such as HIV work or terminal care. We caution, however, against over adjustment for case-mix factors potentially within the control of the family physician.[30]

The calibration of cumulative sums requires a balance between sufficient sensitivity to be able to detect true outliers and the increasing probability of false alarms. We emphasise that here we are referring to statistical false alarms—in other words, any unit whose observed standardised excess mortality results in an alarm despite the true underlying mortality rate for that unit being in control. Several other units may generate signals that are statistical true alarms, but follow-up investigation may reveal a valid explanation. Such units could be seen as medical false alarms. However, these units are one of the subsets that the monitoring system is deliberately designed to detect. Recalling the roles of a monitoring system as suggested by Baker and colleagues,[2] the purpose is not only to detect illegal behaviour (such as Harold Shipman's medical true alarm) but to uncover explanations for other extreme patterns of mortality to provide useful feedback towards improving overall quality of care.

Each of the pilot health authorities were notified of the 11 family physicians who, in addition to Shipman, signalled at K=4 and h=3. However it is too early to gain formal feedback from them. We believe that almost certainly the raised mortality rates in most, if not all, the units will be explained through case mix, data quality issues, or both—in other words that these units represent a mix of statistical and medical false alarms, and not medical true alarms.

Conclusions

We envisage cumulative sum charts being used as a governance tool for monitoring performance since they enable a first-pass analysis of the data and can highlight units with unusual outcomes. We caution however, that the charts cannot by themselves shed light on the reasons for apparent poor performance. Methods of operation at a local level will be required that enable the clinical explanations for outliers to be readily identified. Even in this context, use of cumulative sum charts, or any measurement of individual or institutional performance, will require a paradigm shift in attitudes to performance monitoring among health-care professionals and by the public and media, who will be asked to put their trust in such monitoring. As others have also argued,[2] performance monitoring should be used as a starting point for an audit and learning opportunity rather than recrimination. In a primary care context, we suggest that local health organisations such as primary care Trusts in the UK would be ideally placed to monitor and act on any surveillance information provided to them, especially in the form of cumulative sum charts. Further investigation of signalling units might be done through audit, improved by the collection of additional information on the death certificate and availability of the electronic health record. If cumulative sum (or any other SPC) charts are to be used in surveillance of family physician mortality, there must be a commitment to investigate the reasons behind apparent poor performance.

Finally, if health services are to deliver high-quality, cost-effective care that leads to improved health through guidance, audit, and best practice, it needs high-quality and timely information. Any method, no matter how complex, for analysing and comparing performance will founder if this is unavailable.

Contributors

P Aylin conceived the project. A Bottle did the data extraction. A Bottle, N Best, and C Marshall did the statistical analysis. All investigators helped to draft the paper.

Conflict of interest statement

This study was funded by the Shipman Inquiry. Since completing the work, PA and AB have received funding from Dr Foster Ltd to do analyses of hospital performance.

References

1 Baker R. Harold Shipman's clinical practice, 1974–1998. London: Stationery Office, 2001.
2 Baker R, Jones D, Goldblatt P. Monitoring mortality rates in general practice after Shipman. *BMJ* 2003; **326:** 274–76.
3 The Shipman Inquiry: independent inquiry into the issues arising from the case of Harold Fredrick Shipman. http://www.shipman-inquiry.org.uk/ (accessed July 15, 2003).
4 Curnow RN. Potential contribution of statistical evidence to the inquiry. Bristol: Bristol Royal Infirmary Inquiry, WIT 0361, 1999.
5 Mohammed MA, Cheng KK, Rouse A, Marshall, T. Bristol, Shipman and clinical governance: Shewhart's forgotten lessons. *Lancet* 2001; **357:** 463–67.
6 Adab P, Rouse AM, Mohammed MA, Marshall T. Performance league tables: the NHS deserves better. *BMJ* 2002; **324:** 95–98.
7 Aylin P, Alves B, Best N, et al. Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1984–86: was Bristol an outlier? *Lancet* 2001; **358:** 181–87.
8 Smith AFM, West M, Gordon K, Knapp MS, Trimble MG. Monitoring kidney transplant patients. *Statistician* 1983; **32:** 46–54.
9 Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc* 1996; **159:** 385–443.
10 Page ES. Continuous inspection schemes. *Biometrika* 1954; **41:** 100–15.
11 Whitehead J. The design and analysis of sequential clinical trials. Chichester: Horwood, 1983.
12 Marshall EC, Best NG, Bottle A, Aylin P. Statistical issues in the prospective monitoring of health outcomes at multiple units. *J R Stat Soc* (in press).
13 Sonesson C, Bock D. A review and discussion of statistical issues in public health monitoring. *J R Stat Soc C* 2003; **166:** 5–21.
14 Vanbrackle L, Williamson GD. A study of the average run length characteristics of the national notifiable disease surveillance system. *Stat Med* 1999; **18:** 3309–19.
15 Terje Lie R, Heuch I, Irgens LM. A sequential procedure for surveillance of Down's syndrome. *Stat Med* 1993; **12:** 13–25.
16 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.
17 Morton A, Whitby M, McLaws M-L, et al. The application of statistical process control charts to the detection and monitoring of hospital acquired infections. *J Qual Clin Pract* 2001; **21:** 112–17.
18 Spiegelhalter DJ, Kinsman R, Grigg O, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac survey. *Int J Qual Health Care* 2003;**15:** 7–13.
19 Steiner SH, Cook RJ, Farewell VT, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000; **1:** 441–52.
20 Frisén M. Evaluations of methods for statistical surveillance. *Stat Med* 1992; **11:** 1489–502.
21 Milner RG. Simultaneous statistical inference, 2nd edn. New York: Springer-Verlag, 1981.
22 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 1995; **57:** 289–300.
23 Aylin P, Best N, Bottle A, Marshall C. Monitoring of mortality rates in primary care 2003. http://www.the-shipman-inquiry.org.uk/documentaryday.asp?from=w&day=160 (accessed July 24, 2003).
24 Genovese C, Wasserman L. Operating characteristics and extensions of the FDR procedure. *J R Stat Soc B* 2002; **64:** 499–517.
25 McCullagh P, Nelder J. Generalized linear models, 2nd edn. London: Chapman and Hall, 1989.
26 Rossi G, Lampugnani L, Marchi M. An approximate CUSUM procedure for surveillance of health events. *Stat Med* 1999; **18:** 2111–22.
27 NHS performance indicators. http://www.doh.gov.uk/nhsperformanceindicators/index.htm (accessed July 15, 2003).
28 Christiansen C, Morris C. Improving the statistical approach to health care provider profiling. *Ann Intern Med* 1997; **127:** 764–68.
29 NHS strategic tracing service. http://www.nhsia.nhs.uk/nsts (accessed July 15, 2003).
30 Goldstein, H. Contribution to the discussion of "Spiegelhalter et al. Commissioned analysis of surgical performance using routine data: lessons from the Bristol Inquiry". *J R Stat Soc A* 2000; **165:** 222–23.