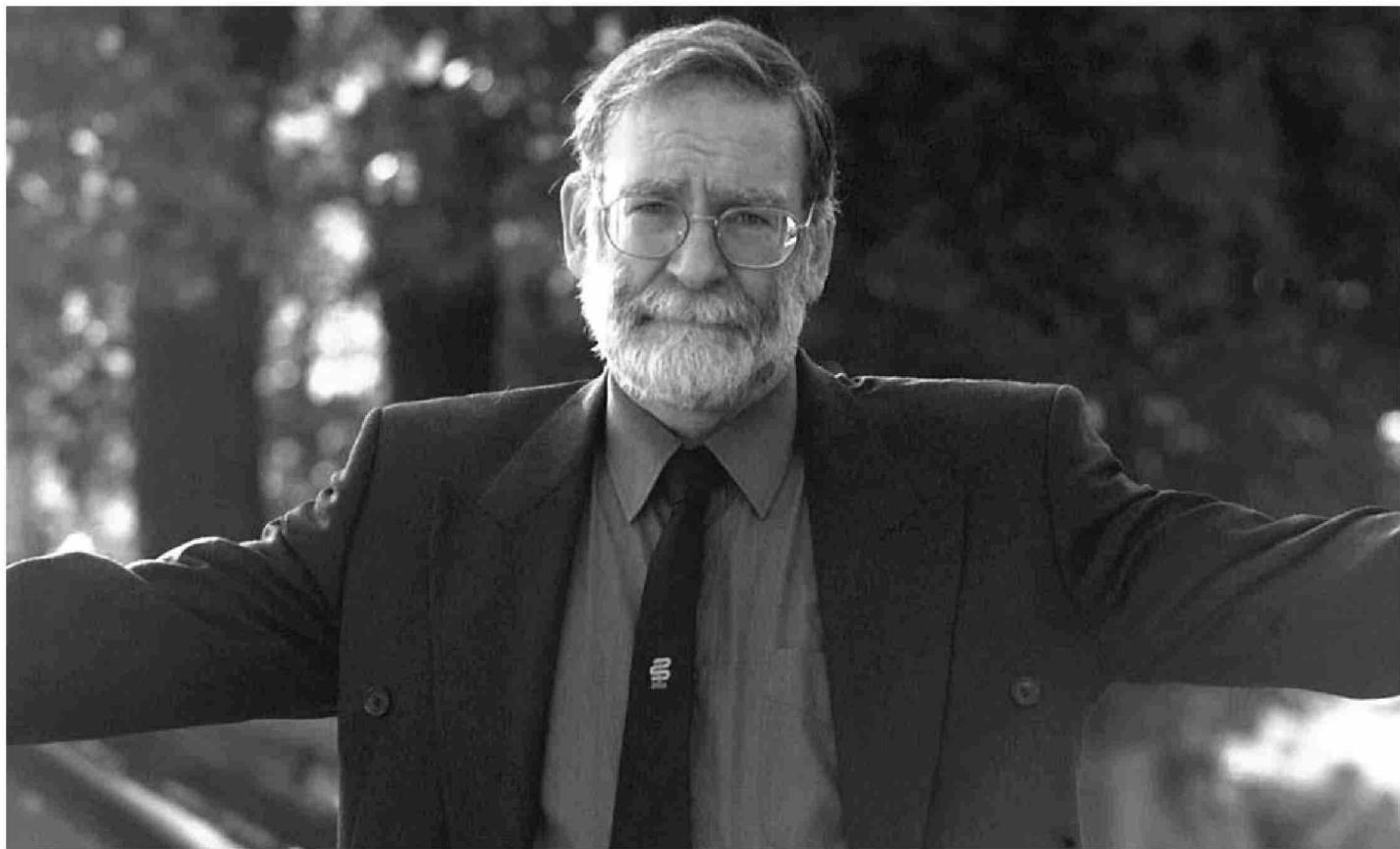


Shipman's statistical legacy

Harold Shipman, who in January committed suicide in prison, has become notorious the world over as one of the most prolific serial killers of all time. His case has also seriously dented public confidence in doctors. **David Spiegelhalter** and **Nicky Best** explain how industrial quality control techniques could be adapted to signal when death rates among a doctor's patients are surprisingly high, and the tricky issues that would arise in implementing such a monitoring system.

Dr Harold Shipman arrives at Ashton-under-Lyne police station (photograph copyright Chris Gleave, MEN syndication)



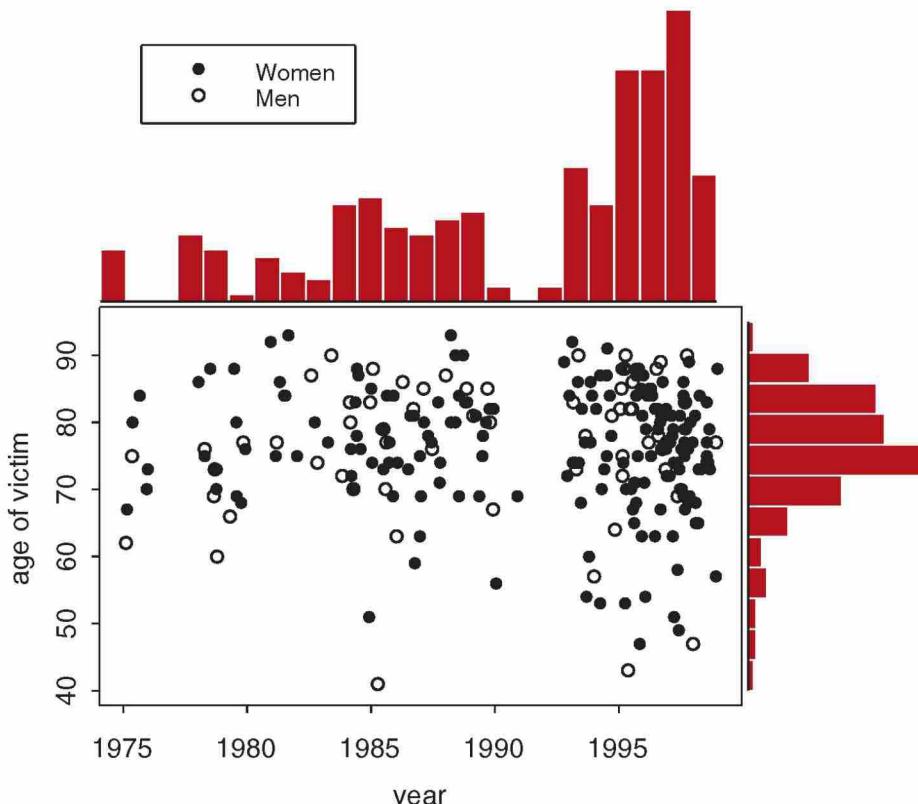


Figure 1. Age and sex of Shipman's victims by date of death

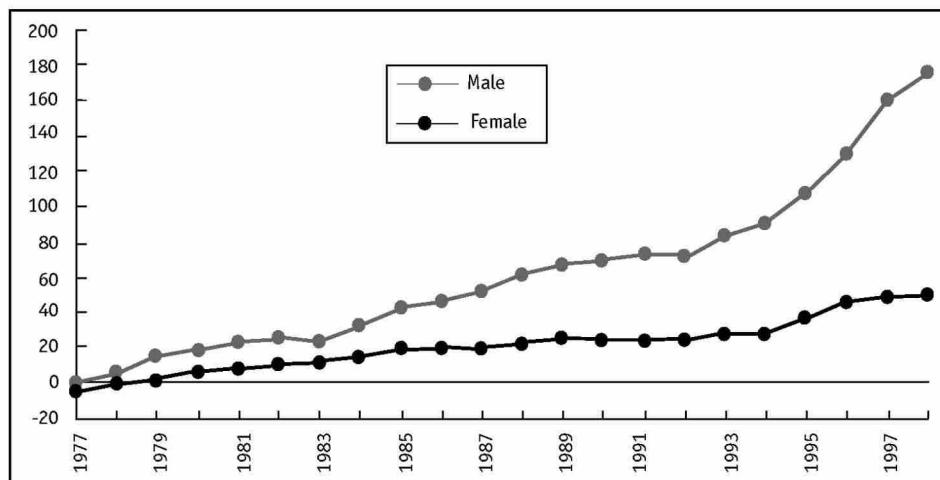
Mrs Laura Macey* was 81 and the cold January of 1986 had brought on her customary bad chest, so it was arranged for her general practitioner (GP), Dr Harold Shipman, to visit her at home. Two of her great-grandchildren called at tea-time and were surprised when the door was not answered. When their aunt looked through the window she saw Mrs Macey lying dead on the sofa, but it was another 16 years before it was concluded that "Shipman must have killed Mrs Macey, almost certainly by lethal injection of diamorphine"¹. She was the 65th of 215 "confirmed" and 45 "probable" victims of Shipman identified by the public inquiry headed by Dame Janet Baker: an inquiry that is still continuing and which is placing considerable reliance on statistical evidence, both in investigating whether Shipman could have been detected earlier, and in making recommendations about future monitoring of GPs' performance.

Shipman's victims tended to be older women (Figure 1), and he clearly went through phases of increased activity, notably after he went into single-handed practice in 1992.

In a careful 2001 report², Professor Richard Baker compared the number of death certificates being signed by Shipman for people

dying in their homes or in his practice, with the number expected from the behaviour of other local GPs. Figure 2 shows that by 1998, the year Shipman was finally arrested, having murdered eight people in March alone, his estimated excess mortality for people aged 65 or over was 175 women and 49 men. This was almost exactly the number of over-64s confirmed to be victims by the Inquiry a year after the Baker report, showing the accuracy of this purely statistical analysis.

Figure 2. Cumulative excess death certificates signed by Shipman, for people older than 64 and who died at home or in his practice



Whistle-blowing

This analysis naturally suggests the question: had such data been available at the time, when might it have been reasonable to have "blown the whistle"? To answer this we need careful statistical analysis that takes into account the caution required when monitoring a process over time, since, if we're not careful, even individuals whose performance is entirely as expected will eventually look odd just by chance if we test them sufficiently often. Back in World War II, statistical quality control procedures were developed for armaments production lines that allowed for this problem of "multiple testing", and these venerable industrial techniques are now being adapted to medical use.

"If we're not careful, even individuals whose performance is entirely as expected will eventually look odd just by chance"

One such technique is the "risk-adjusted cumulative sum" (CUSUM), which sounds complicated but is just a plot of a simple function of the observed and expected deaths against time. If this exceeds a predetermined threshold, chosen to provide the desired trade-off between the chance of falsely detecting a "normal" GP, and the chance of missing a GP with patient mortality rate, say double that expected, then this raises the alarm, signalling that there is something unusual about the GP's mortality rates. A version of this technique suggests³ that Shipman was theoretically detectable at the end of 1985, just before

Mrs Macey was killed, and 13 years before he was finally arrested. But only *theoretically*: in practice data on mortality rates of GPs have never been available.

Signals and noise

In addition, even if it had been possible to use such CUSUM tests to monitor all 27000 or more GPs in England and Wales, we do not know how many entirely innocent GPs might also have "signalled", perhaps because they specialised in terminal care at home, or simply due to chance. Applying such tests to many thousands of GPs at the same time adds a further dimension to the "multiple testing" problem, since the standard alarm thresholds are only designed to balance the probabilities of false and successful detection of a single GP over time.

The Shipman Inquiry therefore commissioned research by the Department of Epidemiology and Public Health at Imperial College London to assess the feasibility of routinely monitoring deaths in general practice⁴. The Imperial team, comprising Paul Aylin, Nicky Best, Clare Marshall and Alex Bottle, considered both the availability of relevant data sources and what statistical methods might be used for this purpose. They found that it was possible to obtain information on the number of patients registered with each GP and the number of these who died each year, by linking various routine datasets collected by the Office for National Statistics and the

National Health Service Information Authority, although relevant data were only available from 1993. From these data the annual mortality rate for each GP could be calculated, but these were found to vary widely (more than 30-fold), even after adjusting for patient age.

Unfortunately, the data contained very little information about patient characteristics that might affect their risk of dying independently of the actions of their GP. This "case mix" information should, ideally, be taken into account when assessing an expected mortality rate for each GP, and therefore an adjustment was included (known as "overdispersion") to allow for the estimated effects of unmeasured case mix. The Imperial team also proposed a way of taking account of multiple testing over thousands of GPs when choosing the alarm thresholds for the CUSUM chart. This involves estimating properties of the monitoring system equivalent to the sensitivity (the proportion of GPs with true excess mortality that are detected) and the "false discovery rate" (the proportion of alarms that are incorrect).

Figure 3 shows the CUSUM charts for the 12 GPs (out of 1009 who were included in the Imperial feasibility study) who signalled as having unusually high mortality rates at some point during the period from 1993 to 1999. Harold Shipman is among these, and his chart is shown in bold (the decline after 1998 corresponds to the locum

who took over his practice). An investigation into the 11 other GPs signalling revealed a substantial proportion of deaths in care homes that was likely to explain their unusual mortality rates.

"Given good quality data, our opinion is that it would be feasible to monitor GP practices using methods such as risk-adjusted CUSUMs"

The Shipman Inquiry has yet to make recommendations on any future monitoring system. Its task is made more complicated by proposals for patients to be registered with practices rather than individual GPs. Given good quality data, our opinion is that it would be feasible to monitor practices using methods such as risk-adjusted CUSUMs. However, since we might expect many "signals" of apparently unusual mortality rates to be due to data quality issues or case mix, these methods should be seen as a "screening tool" for the initial analysis of routine health outcome data that could be used to direct further investigation. It is vital that such investigations then follow a cautious multistage procedure that aims to understand the reasons for a signal and is not just seen as a whistle-blowing exercise.

*The name of Dr Shipman's victim has been changed.

References

1. The Shipman Inquiry. (Available from www.the-shipman-inquiry.org.uk/.)
2. Baker, R. (2001) *Harold Shipman's Clinical Practice 1974–1998: a Review Commissioned by the Chief Medical Officer*. London: Stationery Office.
3. Spiegelhalter, D. J., Kinsman, R., Grigg, O. and Treasure, T. (2003) Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman, and adult cardiac surgery. *International Journal of Quality in Health Care*, 15, 7–13.
4. Aylin, P., Best, N. G., Bottle, A. and Marshall, E. C. (2003) Following Shipman: a pilot system for monitoring mortality rates in primary care. *Lancet*, to be published.

David Spiegelhalter is a Senior Scientist at the MRC Biostatistics Unit in Cambridge. Nicky Best is a Reader in Statistics and Epidemiology, at the Department of Epidemiology and Public Health, Imperial College London. Part of their research work involves statistical methods for detecting divergent performance, and both have given evidence to the Bristol Royal Infirmary and the Shipman Inquiries.

Figure 3. CUSUM charts for the 12 GPs (out of 1009 monitored) who signalled (i.e. whose chart statistic crossed the alarm threshold of 3) between 1993 and 1999: Shipman's chart is shown in bold red

