Statistical issues in the prospective monitoring of health outcomes across multiple units

Clare Marshall, Nicky Best, Alex Bottle and Paul Aylin

Imperial College London, UK

[Received February 2003. Revised February 2004]

Summary. Following several recent inquiries in the UK into medical malpractice and failures to deliver appropriate standards of health care, there is pressure to introduce formal monitoring of performance outcomes routinely throughout the National Health Service. Statistical process control (SPC) charts have been widely used to monitor medical outcomes in a variety of contexts and have been specifically advocated for use in clinical governance. However, previous applications of SPC charts in medical monitoring have focused on surveillance of a single process over time. We consider some of the methodological and practical aspects that surround the routine surveillance of health outcomes and, in particular, we focus on two important methodological issues that arise when attempting to extend SPC charts to monitor outcomes at more than one unit simultaneously (where a unit could be, for example, a surgeon, general practitioner or hospital): the need to acknowledge the inevitable between-unit variation in ‘acceptable’ performance outcomes due to the net effect of many small unmeasured sources of variation (e.g. unmeasured case mix and data errors) and the problem of multiple testing over units as well as time. We address the former by using quasi-likelihood estimates of overdispersion, and the latter by using recently developed methods based on estimation of false discovery rates. We present an application of our approach to annual monitoring ‘all-cause’ mortality data between 1995 and 2000 from 169 National Health Service hospital trusts in England and Wales.

Keywords: Clinical governance; False discovery rates; Hospital episode statistics; Overdispersion; Routine surveillance; Statistical process control charts

1. Introduction

Recent high profile failings in the delivery of health care (Bristol Royal Infirmary Inquiry, 2000, 2001; Horton, 2001; Ramsay, 2000; Secretary of State for Health, 1997) have encouraged calls for the more rigorous surveillance of outcomes such as mortality rates. Several national initiatives have been instituted to facilitate performance monitoring (National Health Service Executive, 1999; Department of Health, 2000) including the recently announced Office for Information on Health Care Performance (Department of Health, 2002).

For any monitoring exercise, there needs to be clarity about its precise objectives. A crucial distinction is whether the aim is to investigate unusual performance (of a unit or individual) within a particular preselected clinical speciality, typified by the comprehensive analysis that was commissioned by the Bristol Royal Infirmary Inquiry (Aylin et al., 2001; Spiegelhalter et al., 2002) or, alternatively, to carry out prospective surveillance of health statistics to detect areas of potential concern about clinical performance. Here we focus on the latter and consider some of the methodological and practical issues surrounding the routine monitoring of health outcome data. We emphasize that our proposed approach could form one part of a performance
‘screening’ tool for the first-pass analysis of routine health outcome data that could then be used to direct further investigation and to target clinical audit.

Outcome rates for a given unit (e.g. surgeon, general practitioner, hospital or health authority) can be viewed as realizations of an underlying process. From a statistical perspective, the aim of any surveillance system is to detect important changes in this process as soon as possible after they have occurred. Methods that have been applied in the context of surveillance include the Kalman filter (Smith et al., 1983), Bayesian changepoint methods (Carter and Blight, 1981) and standard hypothesis testing procedures that apply the classical type I (i.e. false positive) error rates to determine the statistical significance of the observed and expected outcomes. The major problem with the last is that it is a ‘one-off’ test that is carried out only after all the data have been collected. In contrast, the distinctive feature of a prospective monitoring system is that data are accumulating over time and the analysis is repeated at every time point. Statistical process control (SPC) charts address the resulting multiple-testing problem and are among the most widely used methods for such sequential analysis, not least because of their ease of implementation and simplicity of presentation.

Although SPC charts were developed within the context of industrial process control, several researchers have proposed their use to monitor health outcomes. Mohammed et al. (2001) advocated using Shewhart charts in clinical governance, whereas Lovegrove et al. (1997) and Poloniecki et al. (1998) both applied a type of cumulative sum (CUSUM) control chart that plots the cumulative difference between the observed and expected deaths adjusted for case mix using a preoperative risk score, to monitor death-rates following cardiac surgery. Applications of SPC charts in public health surveillance include the detection of outbreaks of various notifiable diseases in the USA (Vanbrackle and Williamson, 1999) and changes in the birth prevalence of Down’s syndrome in Norway (Lie et al., 1993).

All these references considered the surveillance of a single unit over time. Apart from our own recent work on monitoring mortality in general practice (Aylin et al., 2003) and some limited discussion of the issue by Spiegelhalter et al. (2003), we are unaware of any literature that addresses the problem of simultaneous surveillance of health outcomes over multiple units as well as time points. Some important methodological issues must be addressed to extend the SPC chart approach to monitor simultaneously outcomes at $M > 1$ units (where ‘unit’ could refer, for example, to an individual surgeon or general practitioner, a National Health Service (NHS) hospital trust, a primary care trust or a health authority). In particular are the following issues.

(a) SPC charts require that we define when the process being monitored may be considered to be acceptable, or in control, and when it is deemed to have become unacceptable, or out of control. In the context of monitoring multiple units, this requires an understanding of the different sources of variability that are inherent in the observed outcome data for each unit, and in particular the specification of an appropriate in-control distribution that acknowledges the inevitable between-unit variation (overdispersion) in true but acceptable performance levels.

(b) SPC charts are designed specifically to adjust for multiple testing over time, but a new approach is needed to handle the problem of multiple significance testing over units as well as time.

In Aylin et al. (2003) we presented a practical application of these methods to the monitoring of annual mortality in general practice, whereas in the current paper we provide a more detailed discussion of the statistical and methodological issues. In Section 2, we introduce an application to the routine surveillance of annual ‘all-cause’ mortality rates at 169 NHS hospital trusts using data for the years 1995–2000. In Section 3 we outline the control chart method for monitoring
the performance of a single unit over time. In Sections 4 and 5 respectively we consider the two methodological issues that are associated with monitoring multiple units noted above, and we apply these by using the NHS trust data in Section 6. We conclude in Section 7 with some suggestions for future work and a brief discussion of some of the practical and policy implications of establishing a formal performance monitoring system in the NHS.

2. Motivating example

As an application of the methods that are outlined here, we consider the surveillance of all-cause mortality at 169 NHS trusts between 1995 and 2000. The trusts represent non-specialist centres, predominantly district general hospitals. The trust data were obtained from the hospital episode statistics database that is held by the Dr Foster Unit at Imperial College London.

For each trust we have observed mortality counts summed over 80 diagnoses that together account for 80% of all deaths in hospital (Jarman et al., 1999). Expected numbers were obtained through indirect standardization, adjusting for year, age, sex, diagnosis (three-digit international classification of diseases code) and method of admission (elective or emergency). Because a death that occurs after a patient leaves hospital is not recorded within hospital episode statistics, an adjustment is also made for the length of stay. For simplicity we treat admissions as the unit of analysis, though in fact 19% of the total were readmissions. This may be an issue where some trusts have many readmissions, thus diluting their mortality. However, overall, there was little correlation (−0.14) between readmission rates and crude mortality rates for each trust.

For the purposes of the current paper, we assume that these data were collected prospectively, annually, and that we are interested in detecting important changes (both increases and decreases) in all-cause mortality at any of the 169 trusts.

3. Statistical process control charts for monitoring performance of a single unit over time

Several different types of SPC chart have been proposed in the context of health surveillance (see Sonnesson and Bock (2003) for a recent review), but all share the following key features.

(a) A test statistic $S_t$ that is calculated for the unit at each time $t = 1, 2, \ldots$: this statistic is based on some function of the standardized residual $R_t$ between the observed outcome at time $t$ and that expected under the in-control distribution. $S_t$ may also depend on previous values of the residuals, leading to a cumulative sum.

(b) A predefined threshold $h$: if the test statistic $S_t$ exceeds $h$ 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.

(c) Some measure of the performance of the chart in terms of its ability to detect when the underlying process is truly in and out of control: such ‘performance measures’ take the place of the more familiar type I (false positive) and type II (false negative) error rates in ordinary significance tests.

Here, we have used a normal cumulative sum (CUSUM) chart with test statistic defined as

$$S_t = \max(0, S_{t-1} + R_t - K/2)$$

where $R_t - K/2$ is proportional to the log-likelihood-ratio, $\log\left\{f_1(y_t)/f_0(y_t)\right\}$, in which $f_0$ and $f_1$ denote the distributions of the in- and out-of-control processes respectively and $y_t$ is the observed outcome (e.g. the number of deaths) at time $t$. Here we assume that $f_0 \sim N(\mu, \sigma^2)$ and

Prospective Monitoring of Health Outcomes 543
C. Marshall, N. Best, A. Bottle and P. Aylin

$f_1 \sim N(\mu + K\sigma, \sigma^2)$; thus $K\sigma$ represents the increase in the underlying process mean $\mu$ to be detected by the chart. Note that we may wish to specify two different out-of-control shifts to be detected: one representing a warning rate (e.g. $K = 1$ or $K = 2$) and one representing an alarm rate (e.g. $K = 3$ or $K = 4$). We defer discussion of the calculation of $R_t$, the standardized residuals for the observed outcome at time $t$, and our motivation for assuming these to be normally distributed under the in- and out-of-control processes, until Section 4.1.

Analogously, the test statistic for use in a CUSUM chart that is designed to detect a decrease in outcome rate is given by

$$S_t = \min(0, S_{t-1} + R_t + K/2).$$

Our particular choice of SPC chart is motivated by various desirable properties of the likelihood ratio CUSUM statistic compared with other types of SPC chart. In particular, it has been shown to be optimal (among other types of CUSUM chart) in the sense that it gives the greatest chance of detecting a true change in the outcome measure for a given false positive rate (Moustakides, 1986). Another advantage of the CUSUM test statistic is that it is bounded below by 0, whereas the test statistic in the closely related sequential probability ratio test (Wald, 1947) chart can take positive or negative values. If a process remains in control, the sequential probability ratio test statistic becomes increasingly negative, allowing the unit to build up ‘credit’ for good past performance, helping to mask any subsequent deterioration in performance, at least during the early period after a change. This would not seem to be an attractive feature for a health surveillance system.

Another alternative is the exponentially weighted moving average chart (Hunter, 1986). The exponentially weighted moving average test statistic downweights previous values of the residuals $R_1, \ldots, R_{t-1}$ by an exponential function of time, in contrast with the CUSUM statistic which gives equal weight to all previous values. Although the former may seem appealing, it is not always clear how to choose the downweighting function, and so our preference is for equal weights, although we acknowledge that further empirical research is merited to compare these two charts in the present context.

A simple alternative to CUSUMs that has been advocated for use in clinical performance monitoring is the Shewhart chart (e.g. Mohammed et al. (2001)). However, the Shewhart chart statistic is calculated independently at each monitoring time, using only the most recent observation (or, say, the mean of the observations since the last monitoring time), and so does not make use of all the available information about the unit under study. Although it is quick to respond to rapid large changes in the underlying process, it will be slow to detect small or moderate sustained changes (which are more likely to represent the types of performance shift that we are aiming to identify) and tends to be particularly susceptible to random fluctuations in observed outcome (Vanbrackle and Williamson, 1999; Williamson and Weatherby-Hudson, 1999).

3.1. Setting the alarm threshold and deriving chart performance measures

The threshold $h$ is set by optimizing the trade-off between long times until false alarms and short delay times for true alarms. The usual statistical concepts of type I and type II errors are not appropriate in the context of sequential testing, because these error rates depend on time, and in particular the type I error tends to 1 as time tends to $\infty$. Frisén (1992) has provided a useful review of some alternative measures that can be used to summarize the performance of SPC charts for surveillance, the most common of which is the average run length (the number of time points on average before the test statistic first exceeds the alarm threshold $h$). The average run length when the process is in control is denoted by $\text{ARL}_0$; ideally this should be long, since
in this context a signal would represent a false alarm. The average run length when the process is
out of control is denoted by $ARL_1$ and should be as short as possible since a signal would then
represent a true alarm. These run length distributions will depend on the value of the threshold
parameter $h$ and on the size of the out-of-control shift $K$ that is to be detected. Tabulating values
of $ARL_0$ and $ARL_1$ for various values of $h$ and a specified out-of-control shift $K$ enables an
appropriate value of $h$ to be chosen that reflects the desired trade-off between rapid detection
of a true alarm and the risk of obtaining a false alarm too early.

Although it is standard practice to use $ARL$s to select chart thresholds and to summarize
performance, Frisén (1992) pointed out that the distribution of run lengths is usually markedly
skew, so the average may not be the most appropriate summary. Frisén also criticized $ARL_1$ as
a measure of chart performance because it makes an assumption that the shift from in to out of
control occurred at the same time as the monitoring began. In practice the change is more likely
to occur at some unknown time after the surveillance has started. Other alternative assessment
criteria are proposed in the literature, the most relevant of which are

(a) the probability of a false alarm by time $t$ ($PFA_t$)—an increasing function of time since, the
longer monitoring continues, the greater the chance of a false alarm and

(b) the probability of successful detection within time $t$ of a change ($PSD_t$)—the probability
that the system signals (a true alarm) within a certain time interval after the change from
in control to out of control actually occurred, given that there were no false alarms before
the change. This probability will depend on the time at which the true change to out of
control took place and the size of the change.

Table 1 summarizes some of these performance measures for a CUSUM chart for data or stan-
dardized residuals that are assumed to follow a standard normal distribution when the process is
in control. (Values are based on 100000 simulation runs; Monte Carlo standard errors were also
computed and were negligible.) An examination of these performance measures demonstrates
the trade-off between minimizing the risk of false alarms by raising the alarm threshold and
retaining a good chance of detecting a genuine change. For example, suppose that we wanted to
detect a 1-standard-deviation ($K = 1$) increase in the process mean, and the data were monitored
yearly. The second row of Table 1 shows that choosing a threshold of $h = 3$ gives a chart that has
a 3% chance of falsely signalling within 6 years ($PFA_6$) (where 6 years is the length of the time
series in our motivating example). If the unit is truly out of control (which, in this example,
corresponds to the unit having a mean that is 1 standard deviation above that of the in-control
process) at the start of the monitoring period, there is a 90% chance that the chart will correctly
signal an alarm between the first and 12th year of monitoring (based on the fifth and 95th per-
centiles of the out-of-control run length distribution $RL_1$); if the unit starts off in control, but
becomes out of control at some subsequent time after monitoring starts (assumed to be drawn
uniformly over the discrete interval $[0, 100]$), the chart has a 67% chance of correctly detecting
this change within 6 years of its occurring ($PSD_6$). If a lower threshold is chosen (e.g. $h = 2$; see
the first row of Table 1) then there is a greater chance of correctly detecting an out-of-control
unit within a given time, but at the price of higher false alarm rates. Discrimination between
true negative and true positive results for any given threshold $h$ improves with increasing $K$.

4. Modelling the in-control process: adjusting for multiple sources of within-
and between-unit variation

The first goal of the surveillance exercise is to understand the inevitable variation in performance
within and between units; this then allows us to assess whether the performance outcomes that
<table>
<thead>
<tr>
<th>Shift K</th>
<th>Alarm threshold h</th>
<th>Percentiles of RL₀ distribution</th>
<th>PFA₆ (%)</th>
<th>Percentiles of RL₁ distribution</th>
<th>PSD₆ (%)</th>
<th>PSD₆ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>&gt;100</td>
<td>12.2</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>9</td>
<td>&gt;100</td>
<td>3.0</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>21</td>
<td>&gt;100</td>
<td>0.7</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>15</td>
<td>&gt;100</td>
<td>2.0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>98</td>
<td>&gt;100</td>
<td>0.2</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>0.02</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>0.24</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>0.01</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>&gt;100</td>
<td>&gt;100</td>
<td>&lt;0.01</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>

†K, the size of the out-of-control shift to be detected (measured in multiples of the standard deviation of the in-control process); h, the threshold value of the chart statistic that will generate an alarm; PFA₆, the estimated probability of a false alarm within t years; PSD₆, the estimated probability of a true alarm within t years of a change from in to out of control; RL₀, the estimated in-control run lengths (the number of time points from the start of monitoring until a false alarm); RL₁, the estimated out-of-control run lengths (the number of time points from the start of monitoring until a true alarm signals for a process that was out of control from the start).

are observed in the unit of interest can reasonably be assumed to fit into the pattern of variation that is seen for the other units. Variability can be due to the following.

(a) Random (chance) fluctuations: we shall refer to this as type A variation.

(b) Known sources of variation (type B), e.g. differences across units, or over time, in the age and sex distribution or level of socioeconomic deprivation of the patient populations: these are often referred to as patient ‘case mix’.

(c) Inevitable between-unit variability due to the net effect of many small, unmeasured factors (type C1): these are factors that relate to, for example, the quality of the data and further unmeasured differences in patient case mix. The mortality rates for most units are likely to be influenced by such factors; although the individual effect of any particular factor may be small, the combined effect of many small unmeasured factors may lead to the rates for some units appearing unusual. Sustained changes in these factors across all units over time could also lead to, for example, persistent temporal trends in mortality rates.

(d) More systematic unknown sources of variation that are only associated with one or a handful of units (type C2), but which may also lead to unusual mortality rates for those units: exceptionally good or bad performance due to systematic factors within the control of the unit would fall into this category, as would the effects of large unmeasured case mix factors such as the presence of nursing homes or hospices in the unit’s catchment area.

(e) The effect of seasonal variation and extreme external forces (type D): for example, seasonal variations in temperature are known to influence mortality, and epidemics or particularly severe winters may act to increase annual mortality rates for some or all units.
4.1. Adjusting for type A, B and D variation

We shall usually want to adjust the expected (in-control) outcome rate of each unit for variation due to chance (type A), known risk factors (type B) and any observed temporal trends (type D) common to all units. This may be achieved by defining an appropriate sampling distribution and accounting for the effects of known risk factors and common seasonal trends by covariate adjustment using regression or standardization. A major difficulty of adjusting for case mix (known risk factors) is that, unless a clinically acceptable risk adjustment scheme already exists, there may be concerns about the validity of the particular method of risk adjustment that is used for performance analysis. For example, if the risk adjustment scheme is estimated from the data that are being monitored, there is a danger of overcompensation for observed differences in performance between units (Goldstein, 2002).

The choice of reference rates for indirect standardization also requires careful consideration. For example, if the units being monitored cover a wide geographical area, then using national reference rates will simply lead to the detection of broad scale regional differences in the standardized mortality or incidence ratios that are already well established in the epidemiological literature. Local reference rates may therefore be more appropriate for detecting performance variations between individual units, provided that they are sufficiently robust to be assumed known without error.

Strictly speaking, the reference rates that are used to estimate the expected mortality for each unit should be estimated excluding the data for that unit, since if the unit’s mortality rate is actually out of control then this may bias the estimated in-control rate. However, the influence of any one unit is likely to be minimal relative to the total amount of data pooled over all units regionally or nationally, and so for practical reasons it is simpler to calculate a single set of reference rates based on data from all units, rather than to attempt a full ‘cross-validatory’ approach leaving out data for each unit in turn to estimate unit-specific reference rates.

Most SPC charts also assume stationarity of the mean and variance of the in-control process and independence of the rates over time, which is clearly not tenable for most health outcomes, where changes over time in, for example, patient case mix inevitably lead to a non-constant expected outcome rate. This problem can easily be addressed by monitoring standardized residuals \( R_t \) over time \( (t = 1, 2, \ldots) \) as opposed to the observed outcomes—making an initial transformation to normality if the outcome data are counts. It should be emphasized, however, that statistical comparisons of performance based on such approximations may be subject to error if the assumptions of the approximation (e.g. the requirement for large populations at risk) are not adequately met (Christiansen and Morris, 1997).

Within the context of prospective monitoring, Lovegrove et al. (1997) discussed surveillance of mortality following adult cardiac surgery. They assumed a Bernoulli sampling distribution (type A variation) for the outcome of each operation and used the Parsonnet scoring system (Parsonnet et al., 1989) to obtain expected probabilities of death after adjustment for preoperative risk (type B variation). Rossi et al. (1999) monitored respiratory mortality in males in North Tuscany, assuming a normal approximation to the Poisson sampling distribution for the monthly counts of respiratory deaths. They carried out risk adjustment by calculating expected counts via indirect standardization using age-specific mean monthly rates of a preceding base-line period. More complex temporal patterns exhibited by the underlying in-control process, such as regular seasonal fluctuations or temporal correlation, may be modelled by using time series methods (e.g. Williamson and Weatherby-Hudson (1999)).

In our analysis of the hospital trust mortality data, we adjust for known case mix variables
and common temporal trends via standardization and assume a normal approximation to the Poisson sampling distribution (see Section 6 for further details).

4.2. Adjusting for type C sources of variation

The effect of unknown or unmeasurable risk factors (type C variation) is likely to lead to greater variability in the observed outcome rates between units and over time than would be expected due to random binomial or Poisson variation. If we are interested in detecting differences between units in outcomes due to the combined effects of these factors, then the appropriate binomial or Poisson model, or a normal approximation, is reasonable (having first adjusted for known risk factors and seasonal variation as already discussed). However, if we are primarily interested in detecting those units with large systematic factors (type C2 variation) leading to unusual (out-of-control) outcome rates, then ideally we should also adjust the in-control outcome rates for the type C1 component of variation due to the combined effect of many small unmeasured factors. This can be thought of as an adjustment for overdispersion; if it is not made, then too many false alarms will result.

Our approach to this problem is to inflate the in-control variance by an estimate of the level of overdispersion (Farrington et al., 1996). This method effectively allows the observed data to be more variable than assumed under a Poisson sampling distribution (or its normal approximation) while still being regarded as arising from an in-control process. The net effect is that the transformed residuals that are used to form the chart statistic for monitoring are smaller (less extreme) than would be the case if only type A, B and D sources of variation were allowed for.

One limitation of this approach lies with the estimation of the overdispersion factor. What is required is an estimate of the amount of overdispersion (extra-Poisson variation) in mortality counts arising from an in-control process. In the example in Section 6, the only data that were available to estimate this were the actual mortality counts being used for surveillance, and some of these are likely to be realizations from units that are actually out of control. Any estimate of overdispersion that is based on these data is therefore likely to overestimate the true amount of overdispersion since the out-of-control observations will tend to make the data more variable than would be the case if they all arose from an in-control process. Since we had no way of selecting only the in-control data before carrying out the surveillance, we had no option but to estimate the overdispersion factor based on all the data. However, we obtained separate estimates for each year and chose the median overdispersion estimate in an ad hoc attempt to compensate partially for possible overinflation of the estimates. Furthermore, if the proportion of true out-of-control units in the data is small, then the overinflation of the overdispersion estimate should be small.

In any future application of this methodology, it should be possible to improve on this overdispersion estimate by using historical data on a subset of units that are known to have acceptable performance levels. However, for illustration in the present example, we are treating the historical data as if they were being collected prospectively.

5. Acknowledging multiple testing over units as well as over time

The various measures of chart performance discussed above relate to a single chart for monitoring one process over time and are designed to give a chart with the desired trade-off between successful detection and false alarms. In the context of routine monitoring of mortality data, it is necessary to extend these performance measures to describe the expected rates of false alarms and successful detections among the \( M > 1 \) processes (units) being monitored.

The usual approach to multiple significance testing is to control the overall type I (false
alarm) rate. For example, the widely used Bonferroni correction (Milner, 1981) ensures that if \( M \) significance tests are carried out and all \( M \) null hypotheses are true then the probability of falsely rejecting at least one null hypothesis is less than or equal to the specified overall type I error rate. In the context of monitoring performance of many doctors or health care providers by using sequential probability ratio control charts, Spiegelhalter et al. (2003) suggested setting more stringent alarm thresholds by using a Bonferroni-like adjustment for the number of units being monitored. However, there are various difficulties with this approach in the present context. Firstly, Bonferroni-type adjustments are concerned only with the overall type I error rate and not the power of the multiple-testing procedure to reject false null hypotheses correctly. In the present context, we wish to optimize the trade-off between the false and successful detection rates. Furthermore, as noted in the case of monitoring a single unit, the type I error rate for a sequential analysis is not constant but increases with the length of the surveillance period. Finally, and perhaps most crucially, we are not so much concerned with controlling the probability of obtaining at least one false alarm out of the \( M \) units being monitored as estimating the proportion of all alarms detected that are false—the latter has been termed the false discovery rate (FDR). To our knowledge, there is nothing in the literature that addresses all these issues in the context of multiple sequential analysis tests. This is important to quantify the performance of the proposed SPC charts for monitoring multiple units and to make an informed choice of chart parameters such as the alarm threshold \( h \). In the remainder of this section we present the results of some preliminary calculations that we have carried out to address this issue. In our estimation of the FDR we draw heavily on approaches that were first proposed by Benjamini and Hochberg (1995) and more recently extended by Storey (2002, 2003) and Storey et al. (2004). Although developed for quantifying the performance of multiple non-sequential significance tests, their ideas are as compelling in the current context.

Suppose that \( M \) independent units have been monitored for a specified period of time (say \( t \) time intervals) by using a particular SPC chart, and assume that \( m_1 \) of these have been truly out of control since the start of monitoring, with the other \( m_0 = M - m_1 \) remaining truly in control. Table 2 shows the possible outcomes for these \( M \) units after \( t \) time intervals.

In particular, \( A_t(h, K) \) denotes the total number of alarms by time \( t \), and \( F_t(h, K) \) denotes the total number of false alarms by this time using an alarm threshold of \( h \) for an SPC chart designed to detect an out-of-control shift of \( K \) standard deviations. For notational convenience, we suppress the dependence of these variables on \( h \) and \( K \) in what follows. Following Storey (2002, 2003), we may define the (positive) FDR by time \( t \) as

\[
\text{FDR}_t = E\left[ \frac{F_t}{A_t} \bigg| A_t > 0 \right].
\]

<table>
<thead>
<tr>
<th>Number not signalling an alarm by time ( t )</th>
<th>Number signalling an alarm by time ( t )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of units truly in control</td>
<td>( m_0(K) - F_t(h, K) )</td>
<td>( F_t(h, K) )</td>
</tr>
<tr>
<td>Number of units truly out of control</td>
<td>( m_1(K) - T_t(h, K) )</td>
<td>( T_t(h, K) )</td>
</tr>
<tr>
<td>Total</td>
<td>( M - A_t(h, K) )</td>
<td>( A_t(h, K) )</td>
</tr>
</tbody>
</table>

Table 2. Possible outcomes after running SPC charts on \( M \) independent units for a fixed surveillance period of length \( t \).
Note that the more obvious definition of an FDR, $E[F_t|A_t]$, is problematic since there is a positive probability that $A_t = 0$. Under the assumption that the chart test statistics for each unit are independent and identically distributed random variables (in our case, arising from a mixture distribution of the in- and out-of-control distributions), Storey (2003) showed that equation (1) is equivalent to

$$FDR_t = \Pr(\text{unit truly in control} | \text{alarm by time } t)$$

$$= p_0 \Pr(\text{alarm by time } t | \text{unit truly in control}) \{ p_0 \Pr(\text{alarm by time } t | \text{unit truly in control}) + (1 - p_0) \Pr(\text{alarm by time } t | \text{unit truly out of control}) \}^{-1} \quad (1)$$

where $p_0 = m_0/M$, the proportion of units that are truly in control (see Appendix A for a proof). Hence the FDR is in some sense similar to 1 minus the positive predictive value of a standard diagnostic testing procedure. The probabilities in the numerator and right-hand side of the denominator in equation (1) are respectively the probability of a false alarm ($PFA_t$) and the probability of successful detection ($PSD_t$) by time $t$ for a single chart, as discussed in Section 3.1. We use the Monte Carlo estimates of these quantities (e.g. Table 1) to estimate the FDR, conditionally on fixed $p_0$, by

$$\hat{FDR}_t | p_0 = \frac{p_0 \hat{PFA}_t}{p_0 \hat{PFA}_t + (1 - p_0) \hat{PSD}_t}.$$

Another useful quantity to consider is the successful detection rate by time $t$ (i.e. the proportion of all out-of-control units that signal an alarm by time $t$, given that at least one such unit signals). This is analogous to the sensitivity of a standard diagnostic test and is defined as

$$SDR_t = E \left[ \frac{T_t}{m_1} \mid m_1 > 0 \right]. \quad (2)$$

Under the assumption that the chart test statistics for each unit are independent and identically distributed random variables, and that all out-of-control units have been so since the start of monitoring, equation (2) can be shown to be equivalent to the probability of successful detection for a single chart, $PSD_t$ (see Appendix A for a proof). Hence we use the Monte Carlo estimates of the latter (e.g. Table 1) to obtain

$$\hat{SDR}_t = \hat{PSD}_t.$$

Following a similar line of reasoning, it would also be possible to define quantities that are analogous to 1 minus the negative predictive value—termed the false non-discovery rate by Genovese and Wasserman (2002) and Storey (2003)—and to the specificity of the monitoring system. However, here we focus only on the FDR and the sensitivity SDR of our system, since we envisage it being used primarily as a ‘screening’ tool for the identification of units that might trigger further investigation into the quality of the data, case mix, organizational issues and quality of care. In this context, our primary concern is to design a monitoring system whereby the units that signal have a high probability of being truly out of control. In a situation where the primary concern was to avoid failing to identify a true out-of-control unit (e.g. failing to detect another Harold Shipman or Bristol Royal Infirmary incident), then the specificity and false non-discovery rate of the system would be of interest.

Note that neither FDR$_t$ nor SDR$_t$ depends on the total number of units being monitored (in contrast with methods such as the Bonferroni correction), but FDR$_t$ does depend on the
Fig. 1. Multiple chart performance measures for monitoring $M > 1$ units for six time points (years), shown as a function of the proportion of units that are truly out of control: (a) $K = 1, h = 2$; (b) $K = 1, h = 3$; (c) $K = 2, h = 2$; (d) $K = 2, h = 3$; (e) $K = 2, h = 4$; (f) $K = 3, h = 2$; (g) $K = 3, h = 3$; (h) $K = 3, h = 4$.

The proportion of those units that are truly out of control. Both FDR and SDR depend on the length of the surveillance period over which the chart performance is being evaluated and will increase as $t$ increases. Fig. 1 shows estimates of these chart performance measures for monitoring $M > 1$ units for $t = 6$ time points (e.g. 6 years) for a range of fixed values of $p_1 = 1 - p_0$, the true proportion of units that are out of control. Figs 1(a)–1(c) show how FDR decreases as $p_1$ increases (note that we would typically expect a relatively small proportion of units to be truly out of control in the context of monitoring mortality data, so attention should be focused on the FDRs that are shown in the left-hand portion of each plot); Figs 1(d)–1(f) show SDR, which remains constant irrespective of the number of true out-of-control units. The three different plots in each row correspond to CUSUM charts designed to detect different levels of out-of-control performance ($|K| = 1, 2, 3$); these show that, as the out-of-control shift to be detected becomes more extreme, the charts are better able to distinguish correctly between true in- and out-of-control units. The three curves in each plot correspond to charts with alarm thresholds of $|h|$ either 2, 3 or 4; comparing the top and bottom plots in each ‘column’ illustrates how increasing the alarm threshold decreases the FDR, but at the expense of a correspondingly
Table 3. Performance measures for $M > 1$ CUSUM charts assuming standard normal in-control processes and $N(K, 1)$ out-of-control processes†

| Magnitude of shift to be detected, $|K|$ | Magnitude of threshold, $|h|$ | Multiple chart performance measure $\hat{SDR}_6$ (%) $\hat{FDR}_6 | p_0$ (%) for the following numbers of units truly out of control‡: |
|---|---|---|---|---|---|---|
| 1 | 2 | 83.7 | 82.4 | 69.5 | 51.7 | 39.9 | 31.6 |
| | 3 | 67.3 | 59.2 | 41.3 | 24.8 | 17.0 | 12.5 |
| | 4 | 48.6 | 30.8 | 17.7 | 9.2 | 5.9 | 4.2 |
| 2 | 2 | 93.7 | 40.4 | 24.7 | 13.3 | 8.7 | 6.2 |
| | 3 | 82.6 | 8.6 | 4.4 | 2.1 | 1.3 | 0.9 |
| | 4 | 67.8 | 5.1 | 1.6 | 0.3 | 0.2 | 0.1 |
| 3 | 2 | 96.5 | 0.7 | 0.4 | 0.2 | 0.1 | 0.08 |
| | 3 | 89.0 | 0.4 | 0.2 | 0.1 | 0.05 | 0.04 |
| | 4 | 78.5 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |

†False alarm rates are given under the assumption that 3%, 6%, 12%, 18% and 24% of units are truly out of control—corresponding, in our application, to five, 10, 20, 30 and 40 trusts respectively. $K$, the size of the out-of-control shift to be detected (measured in multiples of the standard deviation of the in-control process); $h$, the threshold value of the chart statistic that will generate an alarm; $\hat{FDR}_6 | p_0$, the estimated percentage of alarms observed in 6 years that are false, given that $100(1 - p_0)/%$ of units are truly out of control; $\hat{SDR}_6$, the estimated percentage of out-of-control units that are successfully detected within 6 years.

‡The proportion out of 169 equals $100(1 - p_0)/%$.

lower successful detection rate. Numerical values for $\hat{FDR}_t$ and $\hat{SDR}_t$ for selected combinations of $|K|, |h|$ and $p_1$ are also summarized in Table 3. Note that these chart performance measures apply both to charts designed to detect an increase in outcome and to those designed to detect a decrease.

5.1. Estimating the number of in-control units

As part of his work on the estimation of FDRs, Storey (2002) proposed a method for estimating $m_0$, the number of true null hypotheses. In the present context, $m_0$ corresponds to the number of true in-control units, and we outline here how Storey’s method may be used to estimate this. Suppose that we have monitored $M$ units for a fixed surveillance period of $t$ time points, and let $T_{ni}^{\max}$ denote the maximum value of the CUSUM statistic for the $i$th unit ($i = 1, \ldots, M$) over this time. Storey’s method makes use of the fact that, conditional on having correctly specified the in-control distribution, a high proportion of the maximum CUSUM statistics that are less than some ‘well-chosen’ threshold $\lambda$ would be associated with in-control units. Under this assumption, a conservative (i.e. upper) estimate of $p_0 (= m_0/M)$ is

$$\hat{p}_0(\lambda) = \min \left( 1, \frac{\text{observed proportion of test statistics} < \lambda}{\text{expected proportion of test statistics from the in-control distribution} < \lambda} \right)$$

$$\approx \min \left\{ 1, \frac{\sum_{i=1}^{M} I(T_{ni}^{\max} < \lambda)/M}{1 - \hat{PFA}_t(\lambda)} \right\}$$

where $\hat{PFA}_t(\lambda)$ is the estimated probability of signalling a false alarm within $t$ time points given an alarm threshold $h = \lambda$ (obtained via Monte Carlo simulation as in Table 1) and hence
corresponds to the expected proportion of maximum CUSUM statistics that exceed \( \lambda \) for units that are in control. The problem remains to choose an appropriate value for \( \lambda \). Storey (2002) suggested a way of optimizing \( \lambda \) based on minimizing the mean-squared error of the resulting estimate of \( \hat{p}_0 \) by using bootstrapping methods, and he presented simulation results that suggest that \( \hat{p}_0 \) is always conservative and very close to the truth.

6. Application to monitoring the performance of 169 National Health Service hospital trusts

6.1. Specifying the in- and out-of-control processes

We make the assumption that the observed mortality count \( y_{it} \) in unit \( i \) and year \( t \) follows an overdispersed Poisson distribution with mean \( E_{it} \), where \( E_{it} \) is adjusted for age, sex, diagnosis (three-digit international classification of diseases code), length of stay and method of admission (elective or emergency). Both to satisfy the stationarity conditions of the SPC charts and to acknowledge the extra-Poisson variability that is inherent in the data, we choose to adopt the Z-transformation

\[
R_{it} = \frac{y_{it} - E_{it}}{\sqrt{(cE_{it})}}
\]

where \( c \) is an estimate of the level of overdispersion that is obtained by using, for example, standard quasi-likelihood methods (McCullagh and Nelder, 1989). These standardized residuals \( R_{it} \) are assumed to follow an \( N(0, 1) \) distribution for in-control units. An out-of-control unit is defined to be a unit whose mean standardized residual mortality count is \( K \) standard deviations higher than expected (for some prespecified value of \( K \)), so that the \( R_{it} \) for such units are assumed to follow an \( N(K, 1) \) distribution.

The observed mortality counts across trusts and over time ranged from 112 to 3132, with expected counts ranging from 110.7 to 3361.0. For overdispersed Poisson variates with means in this range, the above transformation provides a good approximation to standard normality.

The year-by-year (quasi-likelihood) estimates of overdispersion for the trust data ranged from 8.6 to 11.2, with a median of 9.7, indicating a high degree of unexplained variability—over nine times what would be expected due to Poisson sampling variability alone. This is typical of the kind of magnitude of the overdispersion that is inherent in routine health data, after adjustment for measured case mix variables.

6.2. Results

The numbers of trusts that signal at any time during the 6-year surveillance period are given in Table 4 for various values of \( K \) and \( h \). The effect of ignoring the overdispersion in the data is clear. At all values of \( K \) and \( h \), many more trusts signal when the overdispersion is not taken into account—in some cases more than five times as many. These extra signals arise from the underestimation of the variance of the in-control distribution.

Focusing on the results of running a CUSUM chart to detect an increase of \( K = 2 \) in standardized residual mortality (\( R_{it} \)), if the alarm threshold is set at \( h = 3 \) then six units (almost 4%) signal. Figs 2(a) and 2(b) show the CUSUM charts for these six units and plots of their standardized residual mortality against time respectively. The residuals are consistently above 0 in all cases and, within 6 years, sufficient evidence accumulates to suggest that the residual mortalities for each of these units cannot be assumed to have arisen from the in-control distribution. Each unit, therefore, signals. Similar results are seen in Figs 2(c) and 2(d) for the 10
Table 4. Percentage of trusts signalling an alarm at any time between 1995 and 2000 for various values of $h$ (the alarm threshold) and $K$ (the increase in standardized residual mortality to be detected)

<table>
<thead>
<tr>
<th>$K$</th>
<th>$h$</th>
<th>Number of trusts signalling (%)</th>
<th>$K$</th>
<th>$h$</th>
<th>Number of trusts signalling (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No adjustment for overdispersion ($c=1$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>87 (51.5)</td>
<td>−1</td>
<td>−3</td>
<td>73 (43.2)</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>76 (45.0)</td>
<td>−1</td>
<td>−4</td>
<td>69 (40.8)</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>71 (42.0)</td>
<td>−2</td>
<td>−3</td>
<td>63 (37.3)</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>62 (36.7)</td>
<td>−2</td>
<td>−4</td>
<td>55 (32.5)</td>
</tr>
<tr>
<td>Adjusted for overdispersion ($c=9.7$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>18 (10.7)</td>
<td>−1</td>
<td>−3</td>
<td>24 (14.2)</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>11 (6.5)</td>
<td>−1</td>
<td>−4</td>
<td>16 (9.5)</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6 (3.6)</td>
<td>−2</td>
<td>−3</td>
<td>10 (5.9)</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2 (1.1)</td>
<td>−2</td>
<td>−4</td>
<td>7 (4.1)</td>
</tr>
</tbody>
</table>

trusts signalling at $h = −3$ on a chart that is designed to detect standardized residual mortality that is $K = 2$ standard deviations below that expected.

Referring to the multiple chart performance measures that are given in Table 3, we would expect around 8.6% or less of our six signals on the $K = 2$, $h = 3$ chart to be false alarms (depending on the true, unknown, proportion of trusts that are out of control). On the basis of this, we can be fairly confident that the majority, if not all, of the signals identify truly out-of-control trusts. It is important to emphasize, however, that, although these ‘true’ signals can be thought of as true ‘statistical’ alarms in the sense that their standardized residual mortality process cannot be assumed to follow the specified in-control distribution, they may be false ‘medical alarms’, in the sense that they could be explained by factors such as errors in the data or case mix rather than quality of care.

We also need to consider how many true out-of-control units may have been missed. Table 3 shows that the $K = 2$, $h = 3$ chart is expected to miss just over 17% of all out-of-control units. These figures tally with the scenario that around seven units were truly out of control and that six of these have been successfully detected within 6 years, with no false alarms. This is in agreement with the estimate of $\hat{m}_0 = 162$ that was obtained by using Storey’s method, which suggests that $169 - 162 = 7$ trusts are truly out of control.

7. Discussion

SPC charts have been widely used to monitor medical outcomes in a variety of contexts and have been specifically advocated for use in clinical governance (Mohammed et al., 2001). However, previous applications of SPC charts in medical monitoring have considered surveillance of a single process over time. Our aim in this paper was to explore some of the methodological implications of attempting to extend SPC charts to monitor health outcomes simultaneously at multiple units over time. Although our findings suggest that such charts could potentially be of value as part of an NHS-wide performance monitoring system, it is clear that their application and interpretation in this context are not straightforward and that some methodological issues remain to be resolved.

A key issue to address is the problem of quantifying the true and false alarm rates of SPC charts, taking into account multiple testing across units as well as over time. In this context, the
estimation of FDRs is more useful than the usual approach of controlling the overall probability of obtaining at least one false alarm (which will rapidly increase to 1 for SPC charts with even moderate sensitivity as the number of units and time points increases). We have shown how the FDR may be estimated for multiple units and a fixed surveillance period. However, the estimates of FDR and SDR that are reported here should not be regarded as definitive. In particular, our calculations are based on the assumption that all out-of-control units have been so since the start of the monitoring period and do not consider the more realistic situation where units become out of control at different times after monitoring has started. Nor do they take account of the problem that specifying a single out-of-control distribution does not fully capture all the patterns of ‘unusual’ performance that may arise; the reality is more likely to be a continuum since trusts will be affected to different degrees by type C1 (unmeasured case mix) sources of variation, plus possibly also type C2 (large, systematic unmeasured factors) sources. The false and successful detection rates also apply only to a fixed length monitoring period; the possibility of extending the concept of in- and out-of-control run length distributions for a single unit to a joint distribution of run lengths for multiple units should, therefore, be explored. Other chart performance measures such as the distribution of run lengths until the first alarm, or the probability that the first unit to signal is truly out of control, also deserve further investigation. Also important is the effect of ‘restarts’ on chart performance estimates—when a chart signals, it may be restarted, either at 0 or at a head start value, typically $h/2$ (Lucas and Crosier, 1982).
Another interesting area for further research would be to explore formal decision theoretic approaches to the choice of the threshold $h$. Bickle (2002), for example, defined a desirability function, which in our notation equates to

$$D(h) = \{1 - \text{FDR}(h)\}b - \text{FDR}(h)c,$$

where $b$ and $c$ denote respectively the benefit of truly detecting an out-of-control unit and the cost of falsely identifying a unit as out of control. He then chose the threshold $h$ that maximizes this function. Genovese and Wasserman (2002) took a similar approach but instead balanced formally the cost of a false alarm with that of missing a true alarm and minimized the resulting loss function.

A second key methodological issue that we have considered is the problem of how to adjust for the inevitable between-unit variation in performance outcomes due to the net effect of many small unmeasured factors, particularly unmeasured case mix. We deliberately distinguish such variation from that due to more systematic factors that could reflect some aspect of the quality of care that is provided. In designing the CUSUM charts that are presented here, the choice of the out-of-control distribution has been motivated primarily by attempts to detect units whose performance may be influenced by one or two major systematic factors (type C2 sources of variation) leading to a large shift in the mean standardized residual mortality. In practice, however, the charts also appear to be detecting units whose residual mortality is systematically different from expected owing to the net effects of unmeasured case mix and other small errors in the data, but who might otherwise be considered ‘acceptable’ if the reasons for this type C1 variation were known. We have attempted to guard against this to some extent by inflating the variance of the in-control distribution to reflect the additional uncertainty in expected mortality for each unit due to these unmeasured factors. If the unmeasured case mix fluctuates randomly from year to year for each trust—for example, in some years a trust might treat patients that tend to be sicker than average, whereas in other years the patients may be less sick—then this should provide a reasonable description of the in-control distribution. However, some specialist trusts may have unmeasured case mix that remains relatively stable over time, inducing temporal dependence in their standardized residual mortalities, and hence they cannot be assumed to arise randomly from the overdispersed marginal in-control distribution (although nor do they necessarily arise from the specified out-of-control distribution). The implication of this is that their CUSUM charts will gradually rise year by year, eventually signalling. Our method, therefore, will detect not only those units with extreme residuals that genuinely arise from the specified out-of-control distribution but also those units whose performance outcomes show small but sustained differences from that expected over time. If such charts are viewed as a screening tool to identify units for further investigation to understand better the causes of performance variations, then this may actually be an attractive feature.

An alternative approach for handling type C1 variation would be to specify a hierarchical model for the outcomes for each unit and time point. This approach explicitly models the between-unit variation that is attributable to type C1 sources, and appropriate distributional assumptions can be made to reflect the expected magnitude and nature of these sources. In effect, such hierarchical models adjust the estimate of the expected mortality count for each unit to reflect true differences between units in the in-control process due to type C1 sources. There are methodological issues that need to be resolved before a hierarchical modelling approach could be used in this context, however. One problem, similar to that of estimating the overdispersion factor, is that the between-unit (random-effects) distribution of in-control outcomes should ideally be estimated only by using data from in-control units, and not the entire data set. Alternatively, if an appropriate risk adjustment model already exists (such as the Parsonnet score for adult
cardiac surgical outcomes (Parsonnet, 1989)) then this could be used, together with information on the expected distribution of risk factors among the population of patients being treated, to predict the distribution of true in-control outcomes across units. The question then arises about how this distribution should be used to help to specify the SPC chart statistic for each unit. One option is to integrate over the random-effects distribution and to assume that the standardized residuals for each unit arise from a common marginal in-control distribution—this is effectively the same as the overdispersion approach that is used in this paper. Alternatively, following a suggestion by Farewell (2002), it might be sensible to predict a random effect for each unit from the upper tail (or lower tail if the aim is to detect good performance) of the distribution and to assume that the unit is in control if its observed outcomes are no more extreme than those which are expected for a unit with such a random effect. This is a conservative approach that corresponds to assuming that units are in control provided that they are performing no worse (or better, if detection of exceptionally good performance is the aim) than would be expected if the unmeasured case mix for that unit was particularly extreme. A third alternative might be to design SPC charts to monitor the estimated unit-specific random effects themselves, rather than the observed data. These suggestions deserved further investigation, although there is a general concern that hierarchical modelling might detract from the simplicity of the basic SPC chart, making it computationally more involved and less straightforward to justify to a non-statistical audience.

This paper has focused on death as an outcome but CUSUMs can equally be applied to process outcomes such as lengths of stay and waiting times. The routine use of CUSUM charts—or indeed any measurement of performance, either of individual clinicians or of whole systems such as hospitals—will require a fundamental change in attitude towards performance assessment. Appropriate presentation will assist the redirection of the focus from recrimination to process understanding and system improvement.

Finally, if the NHS is 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 sophisticated, for monitoring and comparing performance will break down if this is unavailable.

Acknowledgements

The work was partly funded by the Shipman Inquiry. PA and AB are funded by Dr Foster Ltd to conduct analyses into hospital performance.

Appendix A

Below we provide proofs of the results for the FDR, and SDR, stated in Section 5 (based on Storey (2003)).

\[
\text{FDR}_t = E \left[ \frac{F_t}{A_t} \bigg| A_t > 0 \right] = \sum_{k=1}^{M} E \left[ \frac{F_t}{A_t} \bigg| A_t = k \right] Pr(A_t = k | A_t > 0) = \sum_{k=1}^{M} E \left[ \frac{F_t}{k} \bigg| A_t = k \right] Pr(A_t = k | A_t > 0).
\]

Since the chart test statistics are independent across units, it intuitively follows that

\[F_t | A_t = k \sim \text{binomial}\{k, Pr(\text{unit truly in control} | \text{alarm by time } t)\} .\]
Hence
\[
\text{FDR}_t = \frac{\sum_{k=1}^{M} k \Pr(\text{unit truly in control} \mid \text{alarm by time } t) \Pr(A_t = k \mid A_t > 0)}{\sum_{k=1}^{M} k \Pr(A_t = k \mid A_t > 0)} = \Pr(\text{unit truly in control} \mid \text{alarm by time } t)
\]
and result (1) follows by application of Bayes theorem.

The proof that \( \text{SDR}_t = \text{PSD}_t \) follows in the same way, and by noting that since the chart test statistics are independent then
\[
T_t \mid m_1 = k \sim \text{binomial}\{k, \Pr(\text{alarm by time } t \mid \text{unit truly out control})\}.
\]

If we further assume that all out-of-control units have been so since the start of monitoring, then \( \Pr(\text{alarm by time } t \mid \text{unit truly out control}) \) is equal to the probability of successful detection by time \( t \) for a single unit, i.e. \( \text{PSD}_t \). The result \( \text{SDR}_t = \text{PSD}_t \) follows easily.

References


