Journal of the Royal Statistical Society
SERIES A Statistics in Society

J. R. Statist. Soc. A (2012)
175, *Part* 1, *pp.* 1–47

Statistical methods for healthcare regulation: rating, screening and surveillance

David Spiegelhalter,

Medical Research Council Biostatistics Unit, Cambridge, UK

Christopher Sherlaw-Johnson,

Care Quality Commission, London, UK

Martin Bardsley and Ian Blunt,

Nuffield Trust, London, UK

Christopher Wood

Care Quality Commission, London, UK

and Olivia Grigg

University of Lancaster, UK

[*Read before* The Royal Statistical Society on *Wednesday, June 22nd, 2011, the President*, Professor V. S. Isham, *in the Chair*]

Summary. Current demand for accountability and efficiency of healthcare organizations, combined with the greater availability of routine data on clinical care and outcomes, has led to an increased focus on statistical methods in healthcare regulation. We consider three different regulatory functions in which statistical analysis plays a vital role: rating organizations, deciding whom to inspect and continuous surveillance for arising problems. A common approach to data standardization based on (possibly overdispersed) Z-scores is proposed, although specific tools are used for assessing performance against a target, combining indicators when screening for inspection, and continuous monitoring using risk-adjusted sequential testing procedures. We pay particular attention to the problem of simultaneously monitoring over 200000 indicators for excess mortality, both with respect to the statistical issues surrounding massive multiplicity, and the organizational aspects of dealing with such a complex but high profile process.

Keywords: Funnel plot; League table; Overdispersion; Risk-adjusted cumulative sum; Risk-based inspection; Statistical process control

1. Background

The National Health Service (NHS) is the dominant provider of healthcare in England and relies on funding from national taxation. The traditional model of a regulator of such a public service involved a regime of special data collection combined with regular inspections to place institutions into a category reflecting their degree of 'success' or 'failure'. There have recently been two demands that are changing this model. First, calls for greater efficiency and

Address for correspondence: David Spiegelhalter, Medical Research Council Biostatistics Unit, University Forvie Site, Robinson Way, Cambridge, CB2 0SR, UK.
E-mail: D.Spiegelhalter@statslab.cam.ac.uk

 0964–1998/12/175001

less bureaucracy have led to a more 'risk-based' approach to make the process more targeted and proportionate, both to reduce the data collection and inspection burden on institutions and to optimize the use of regulatory resources. These ideas have grown from industrial applications (Straub and Havbro-Faber, 2005) and were adopted in the UK by bodies such as the Financial Services Authority (2000) and the Tenants Services Authority (2010). Second, recent 'scandals' in the health service have led to expectations of rapid detection of emerging problems which inevitably involves a degree of surveillance. In this paper we describe the statistical aspects of these innovations when implemented by a healthcare regulator charged with rating and inspecting institutions as well as keeping them under continuous surveillance. Although our focus is healthcare we feel that much of the methodology is readily applicable in a wide range of other contexts.

The current healthcare regulator in England is the Care Quality Commission (CQC) which, in April 2009, replaced its predecessor, the Healthcare Commission, under which the methods that are described in this paper were developed. Here we consider the statistical methods that have been adopted for three different contexts: awarding an annual performance rating on each NHS trust, deciding which to inspect and surveillance for early detection of potential problems. Trusts are the legally constituted responsible organizations and include acute and specialist trusts ($n = 169$ in 2008–2009) as the main providers of acute hospital care in the English NHS, whereas primary care trusts ($n = 152$) are currently responsible for the commissioning of primary and secondary healthcare in a geographical area. In addition there are mental health trusts ($n = 57$), ambulance trusts ($n = 11$) and learning disability trusts ($n = 3$). Similar approaches are used to monitor independent sector healthcare organizations but are not discussed here, and social care is also not included.

We emphasize that the requirements and processes of an agency such as the CQC are in a constant state of development, and the precise methods that are described in this paper are not necessarily those that will be used in future regulation duties; in particular, the inspection process is currently being restructured. Specific procedures are therefore largely discussed in the past tense. Nevertheless the generic statistical ideas are expected to remain and are considered in the present tense.

A common theme to all these application areas is the variety of types of data: there may be standardized mortality or incidence ratios, proportions, survey responses, counts of adverse events, categorical data and even qualitative 'intelligence'. Each area requires use of multiple items of data on each trust, each of which can be scored against a standard, which may be either an externally set 'target' or an expected value based on an aggregate performance. In some circumstances these standardized scores need to be aggregated up a hierarchy. Yet, at the same time, there is a demand for methods that are straightforward to implement, can be explained to multiple stakeholders and are robust to potentially mediocre quality data.

The issue of quality of data is crucial. Rather than using the mass of routinely collected data, it may be more efficient to expend resources in collecting fewer, but better quality, items targeted towards determining quality of care and outcomes. However, currently there is strong pressure on a regulator to avoid additional data collection.

Here we discuss three interrelated regulatory strategies. Section 2 concerns indicators and ratings, exploring 'exact' methods for comparing indicators with targets that have contributed to the annual rating given to each trust. Section 3 considers screening to target inspections and introduces the use of Z-scores as a standardized measure of the extent to which an observation is an outlier. This section also covers methods for dealing with overdispersion, aggregation of Z-scores and the selection of trusts for inspection. Section 4 examines methods that are used in continuous surveillance, focusing on the technical and organizational issues arising from

monitoring many data series on large numbers of trusts. Some conclusions are drawn in Section 5, and Appendix A contains specific details for constructing Z-scores.

Though the approaches that are described here were developed within the context of healthcare regulation, they have wider applicability in settings where large amounts of complex operational information are available but need to be summarized at the level of multiple comparable units. We hope that the methods go some way to addressing the criticisms of public sector performance indicators that were identified by the Royal Statistical Society report of 2005 (Bird *et al.*, 2005). Finally, although much of the discussion inevitably focuses on identification of poor performance, the statistical techniques can and should be applied to identifying good performance to emulate.

2. Ratings: the role for 'exact' methods

2.1. Construction of a performance rating

The Healthcare Commission was legally committed to annual publication of ratings for each NHS trust. The CQC delivered a series of annual ratings publications on 392 trusts, the last being in 2008–2009 (Care Quality Commission, 2009a) as this is no longer a legal requirement. Nevertheless the statistical techniques that were adopted as part of the rating system are generally applicable in any situation where performance is to be compared against a target or threshold.

In the 'Existing commitments and national priorities' section of the 2008–2009 ratings, each target was assessed as 'achieved' (3 points), 'underachieved' (2 points) and 'failed' (0 points). Points were then summed and a score obtained by using a set of published thresholds (Care Quality Commission, 2009b): for example if there were 10 indicators, and the total score was under 21 out of a maximum of 30, the existing commitments and national priorities were considered 'not met', and so 'quality of services' was immediately classed as 'weak'. This could happen, for example, with more than three failed indicators. We note that this process can lead to a small change in a single indicator tipping the balance of an overall rating: for example in 2004–2005 Addenbrooke's Hospital in Cambridge dropped from three stars to two stars (under the old star rating system), and careful analysis revealed that this was due to just four too few (out of 417) junior doctors being signed up to the 'New deal' on working hours (Spiegelhalter, 2005a).

Given a prespecified target and an observed performance measure, a definition is needed of what is meant by achieved, underachieved or failed. Many of these were defined by using pre-set thresholds: for example the existing commitment for acute trusts that patients should not wait more than 4 hours in the Accident and Emergency Department was achieved if more than 98% spend less than 4 hours (139 trusts in 2008–2009), underachieved if between 97% and 98% (39 trusts) and failed if less than 97% (12 trusts). However, declaring a trust as 'underachieving' when they fail to meet a specific target can be inappropriate when chance plays a substantial role, say due to fairly low numbers or a rather volatile indicator. This resulted in a degree of statistical tolerance being allowed around the target so that for example in the 2008–2009 assessment 'achievement' for 20 indicators was assessed by 'consistency' with the target. To specify what is meant by 'consistent', we first need to develop a measure of deviation from a standard or target.

2.2. Measuring the deviation of an indicator from a standard or target

We assume an observed indicator y that needs to be compared with a standard t: the standard is assumed known and measured without error. Standards are of two basic types: 'average'

performance across either all trusts or a subset used for 'benchmarking', or an externally set target, possibly based on projections from historical data.

It is traditional to express evidence of deviation from a standard as a P-value, i.e. the probability $P(Y \leqslant y|t,n)$ of an observation at least as extreme as y, were the standard being exactly met (the null hypothesis), where n is an appropriate measure of the precision of measurement, such as sample size. P-values might be transformed to give standard normal deviates or Z-scores—see Section 3. For discrete distributions such as Poisson and binomial, we may use the mid-P-value P_{m} defined as

$$P_{\mathrm{m}}(y,t,n) = P(Y < y|t,n) + P(Y = y|t,n)/2$$

which is a simple form of continuity correction. This quantity is easily obtained from standard software.

If we are to use P-values as the basis for deciding achievement of a target, we need to be able to invert this process to produce critical limits for a prespecified P-value p^*: for each n we need critical (non-integer) values y^* such that $P_{\mathrm{m}}(y^*,t,n) = p^*$. For continuous distributions the value of p^* is obtained by using the inverse cumulative distribution function, but for discrete distributions we require an interpolation procedure. For example, suppose that we want the exact critical limit when Y is Poisson distributed with expected value E under the standard; then we can (Jones *et al.*, 2008)

(a) find y, the lowest integer such that $P(Y \leqslant y) > p^*$,
(b) set $y^* = y - 0.5 + \{p^* - P(Y < y)\}/P(Y = y)$ and,
(c) if $y^* < 0$, set $y^* = 0$.

This guarantees that y^*_{L}, the integer below y^*, has $P_{\mathrm{m}}(y^*_{\mathrm{L}},t,n) < p^*$ and y^*_{U}, the integer above y^*, has $P_{\mathrm{m}}(y^*_{\mathrm{U}},t,n) > p^*$. Hence, although y^* is generally non-integer, it forms an appropriate critical boundary: in fact it is convenient to be non-integer as it is then clear whether an observed count is above or below the critical threshold. A similar procedure can be used for binomial data with target p and denominator n, although with an additional check that, if $y^* > n$, y^* is set to n. These exact critical thresholds can be displayed in funnel plots (see Section 2.3) when a common standard is shared by many trusts, or as pre-set thresholds when each trust has been set its own standard. We emphasize that a high value of the indicator, which almost always indicates 'bad' performance, is associated with a large Z-score and a P-value near 1.

We shall illustrate an application to monitoring performance against targets for methicillin-resistant *Staphylococcus aureus* (MRSA) *bacteraemia* rates in acute trusts. A government objective was declared in 2004 of a 50% reduction in MRSA rates by 2008 and, in consequence, each acute trust was set an explicit target for the number of MRSA cases each year, based on a constant absolute annual reduction corresponding to 20% of a single baseline year (2003–2004): we shall comment on this method for setting trajectories below.

After lengthy discussion the 'underachieve' threshold was set as 1 standard deviation (SD) ($p^* = 0.841$), and 'fail' as 3 SDs ($p^* = 0.999$)—see Section 2.4 for further discussion on these choices. These thresholds were prepublished for different target numbers (Healthcare Commission, 2007). Consider East Lancashire Hospitals NHS Trust (Health Protection Agency, 2010) which in 2003–2004 had 63 cases (Fig. 1) and so was set targets for 50, 37 and 24 cases in 2005–2006, 2006–2007 and 2007–2008 respectively. The underachieve and fail regions are shown in Fig. 1. In fact in two of those years the trust just missed its target but were still counted as achieving since they were within 1 SD.

Thus some tolerance is built into the system, and achievement is described as 'performance consistent with plan'. Of course a problem with allowing tolerances for such a politically sensitive

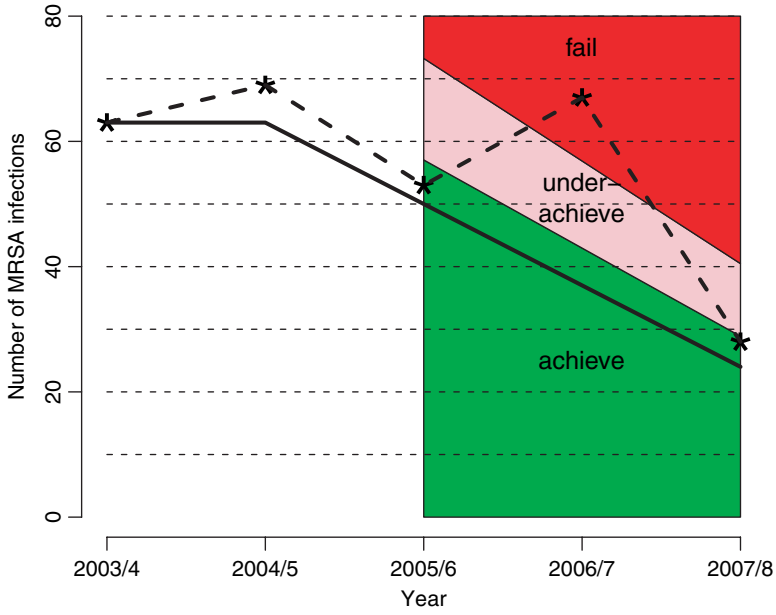

Fig. 1. Critical thresholds for monitoring MRSA counts in East Lancashire Hospitals NHS Trust, with a baseline of 63 and a target absolute annual reduction corresponding to 20% of the baseline (the trust achieved its target in 2005–2006 and 2007–2008 when its actual MRSA count was higher than the target): – – –, observed; ———, target

target is that it may appear that trusts were being let off lightly or, even worse, most trusts might apparently achieve the target and yet national rates not decline appropriately. We have found it helpful to argue that targets should concern the underlying risk that is faced by patients, and the actual number of cases is only an imperfect measure of that underlying risk. In fact, in 2005–2006, 56% of trusts achieved the target, 36% underachieved and 11% failed, whereas in 2007–2008 the proportions were 52% achieved, 39% underachieved and 9% failed. Overall there has been a remarkable fall in MRSA infections, with the national 50% reduction target being met in 2008.

2.3. Funnel plots for changes in standardized mortality or incidence ratios

Suppose that we have observed O and expected E counts in each of two periods, and we assume that the standardized ratio O/E is an estimate of some true underlying standardized mortality ratio SMR. Exact methods based on conditional inference (Breslow and Day (1980), page 93) can be used when assessing a target that $\mathrm{SMR}_2 / \mathrm{SMR}_1 = t$ since, by conditioning on the total observed $O_1 + O_2$, we have under the null hypothesis that the target has been achieved

$$O_2 \sim \text{binomial}\{tE_2/(E_1 + tE_2), O_1 + O_2\}.$$

We can then obtain the critical limits O_2^{crit} by using the methods that were described in Section 2.2, which can be transformed to critical limits for the observed ratio $y = (O_2/E_2)/(O_1/E_1)$, by denoting $O_1^{\text{crit}} = O_1 + O_2 - O_2^{\text{crit}}$, and taking y^{crit} as $(O_2^{\text{crit}}/E_2)/(O_1^{\text{crit}}/E_1)$.

There is an increasing interest in the use of funnel plots as a way of displaying comparative performance data (Spiegelhalter, 2005b), in which the indicator y is plotted against a quantity ρ which is inversely proportional to the null variance s_0^2, so that $\rho = g/s_0^2$ for some constant g. The standard t is drawn as a horizontal line. 'Control limits' are drawn at $t \pm ks_0$, where k

may, for example, be 2 or 3, with the latter corresponding to the classic Shewhart limits. These approximately correspond to P-values of 0.025 (0.975) and 0.001 (0.999) respectively. Once the standard t has been set then the control limits can be 'predrawn' as they do not depend on the data being plotted.

The horizontal scale should be chosen for interpretability: for example $\rho = n$ for proportions and $\rho = E$ for SMRs. For changes we can use a multiple of the inverse sampling variance $1/s^2$ chosen to be on an interpretable scale: for example, when considering the ratio of SMRs, $2/s^2$ is approximately $(O_1 + O_2)/2$, and so the x-axis of a funnel plot against $2/s^2$ could be labelled as approximately representing average observed counts.

Fig. 2 shows an example of the change in MRSA rates between 2006–2007 and 2007–2008 (Health Protection Agency, 2010), testing for differences from the overall change ($t = 0.70$). Out of 168 trusts, 16 (10%) lie outside the central 95% region, compared with 8.4 that we would expect by chance alone, and three lie outside the central 99.8% region. There is clearly a need for caution in interpreting these limits owing to multiple testing, and Jones *et al.* (2008) showed how the funnel limits can be adjusted to control the 'false discovery rate' (FDR)—the proportion of those trusts labelled 'significant' that are expected to be false positive results. Some attention to the issue of multiple comparisons seems essential: for example the US HealthGrades system identifies hospitals as one star or five star if they are significantly better or worse than expected by using a central 90% region, and report that 'approximately 10% to 15% were 1-star hospitals and 10% to 15% were 5-star hospitals' (HealthGrades, 2010). This suggests that between a third and a half of all identified hospitals are falsely positive.

2.4. Discussion

The mapping of performance measures onto overall ratings is a complex and controversial issue which we shall not cover here. However, even at a lower level of aggregation, there are important questions about the choices of how individual targets have been interpreted in terms of

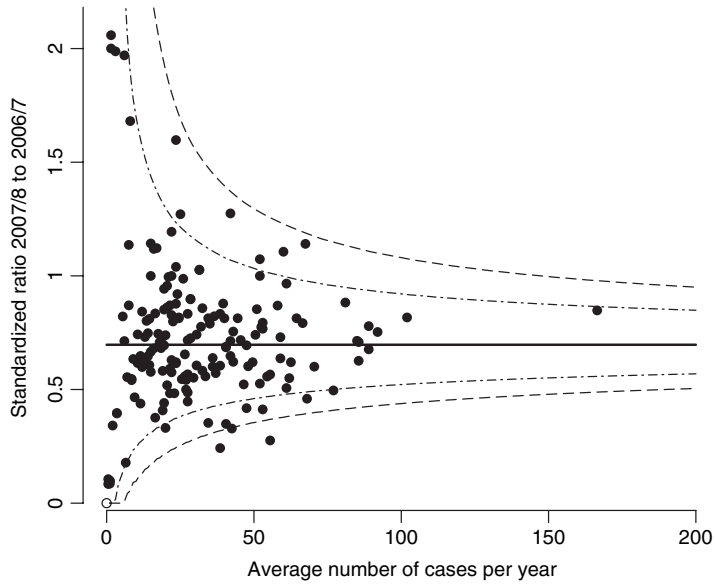

Fig. 2. Funnel plot of the change in standardized MRSA rates from 2006–2007 to 2007–2008 for 168 trusts: – – –, 99.8% limits; · · · · · · ·, 95% limits

sets of specific indicators and the selection of tolerance thresholds. The Department of Health set MRSA reduction targets as part of a process to stimulate rapid improvement in the NHS: an approach that proved successful. Inappropriate statistical methods can, however, lead to unreasonable application of penalties for not meeting targets (Walker *et al.*, 2008).

Setting an individualized target trajectory for trusts requires a baseline, and if that is subject to substantial variability then any measures of improvement will be subject to regression to the mean—see Spiegelhalter (2005c) for a demonstration of this for MRSA data. It is therefore important that a robust baseline be established, either based on a smoothed estimate by using data from a number of periods, or if this is not available then one can use empirical Bayes shrinkage methods to 'adjust' for regression to the mean—see Section 4.5. The MRSA projections were based on a single year and hence trusts who happened by chance to have particularly low rates in 2003–2004 were penalized. For example, Taunton and Somerset NHS Trust had 37 cases in 2002–2003, were sufficiently (un)fortunate to have a low count of 23 in 2003–2004 which became their baseline, and then for the next three years experienced 40, 41 and 38 cases, before finally dropping to 17 in 2007–2008, thus failing their target in three out of four years.

It is not straightforward to decide the appropriate degree of statistical tolerance around a target. A tolerance of 1 SD for underachieve has been used for indicators for which the trust has been set an individual target and there was felt to be considerable potential for influencing performance through appropriate interventions: when the standard is a simple national average or there was seen to be less clear scope for improving performance, greater tolerance has been allowed and the underachieve threshold has been taken as 2 SDs.

The traditional presentation of estimates and confidence intervals is still widely used for comparative performance data such as American College of Surgeons National Surgical Quality Improvement Programme (American College of Surgeons, 2010). However, it has been claimed (Kunadian *et al.*, 2009; Mayer *et al.*, 2009) that funnel plots would be a considerable improvement.

3. Screening for inspection: using a common scale for measuring deviation from a standard

The previous section considered the situation in which the precise value of a single indicator is of primary importance to assess whether a threshold has been breached. We now consider a somewhat contrasting aim that arose when selecting trusts to inspect: summarizing large numbers of indicators to assess more informally the idea of the 'outlyingness' of a trust.

3.1. Introduction to screening

Instead of exhaustive inspection of all trusts on a rolling basis, a 'risk-based' system was introduced by the Healthcare Commission in 2005–2006 to create an inspection regime that was targeted and proportionate. Briefly, the system comprised three stages. First, the board of each trust had to make a public self-declaration about whether they were complying with each of a set of 'core standards' (Department of Health, 2004). Second, each of these declarations was cross-checked against available relevant information to see whether there was a risk of 'undeclared non-compliance': this is the screening process. Third, a proportion of trusts that were deemed most at risk were inspected as well as a random proportion of trusts.

From a statistical perspective we needed, for each standard in turn, to identify relevant data from a wide variety of sources and then to combine the evidence in these data to obtain a measure of 'extremeness' for each trust. The approach recognizes that for some standards there are little or no relevant data on which to base decisions (Bardsley *et al.*, 2009). It is important to

note the shift of emphasis from using indicators for direct judgement (as in the previous section) to using available information simply to direct a set number of annual inspections.

3.2. Z-scores

The basic unit of comparison for all indicators was the Z-score, representing deviation from a standard on a common scale, which allows combination across different types of indicator by using a common set of techniques. It will generally be useful to transform both a 'raw' indicator y and a standard t before conversion to a Z-score: we may also want to transform back at the end for interpretable presentation, which restricts us to easily invertible transformations. For the moment we assume y and t transformed accordingly, and we discuss appropriate transformations later. In what follows we assume that all indicators have been coded so that 'high is bad'.

The 'unadjusted' Z-score is defined as

$$z = (y - t)/s_0 \tag{1}$$

where s_0 is the standard error of y given that the trust exactly meets the standard. Under the null hypothesis that a trust exactly meets the standard, z has mean 0 and SD 1, and if we assume normality then P-values 0.025 (0.975) and 0.001 (0.999) correspond to $z = \mp 1.96$ and $z = \mp 3.10$ respectively, which are very close to 2 and 3 SDs from the standard.

It is important to note that, since we are carrying out a hypothesis test, s_0 is the standard error assuming that the standard is being met and may not necessarily be the same as the reported standard error s that underlies confidence intervals. The difference between these two standard errors explains why, for example, a confidence interval may just include the standard whereas the P-value may indicate that the standard is not being met: the latter is the more appropriate comparison.

It is also vital to emphasize that many indicators will exhibit substantial 'overdispersion', in the sense that the between-trust variability is far higher than would be expected and perhaps a majority of observed unadjusted Z-scores will appear extreme. We shall deal with this issue below.

Details of the calculation of the unadjusted Z-scores for specific types of indicator are shown in Appendix A. The crucial issue in choosing a transformation of y and t is whether, in practice, the bulk of the Z-scores follow a roughly normal distribution, even if overdispersed.

3.3. Overdispersion

'Overdispersion' occurs when the set of trusts exhibits substantially more variability than would be expected from assessments of within-trust sampling error and is clearly beyond a small proportion of 'outliers'.

This behaviour can occur for various reasons. First, indicators that are based on large numbers of cases have a precision that can result in statistically significant differences that are not of practical importance. Second, high overdispersion may reflect indicators that are essentially determined by policy choices rather than events subject to chance, e.g. measures of processes that are largely under the control of the trust. Third, overdispersion can result from grossly inadequate risk adjustment, so that like is not being compared with like. Finally, there will be apparent overdispersion if there are genuine major differences in performance due, perhaps, to variable quality of care.

When using a standard based on 'average' performance, it may then be reasonable to accept as inevitable a degree of between-trust variability in performance and to seek to identify trusts

that deviate from this distribution, rather than deviating from a single standard. In extreme circumstances it may even be better not to use a statistical methodology but simply to assign trusts to bands according to thresholds determined by external judgement. *Underdispersion* may also occur when an indicator is largely under the control of an institution, say in determining resources, and there are clear targets to which all institutions are attempting to adhere closely.

As mentioned in Section 3.2, overdispersion is best handled on a scale in which the distribution of the transformed indicator is reasonably symmetric. The degree of overdispersion then needs to be estimated, but in a way that avoids undue influence of outlying trusts, since these are the very trusts that we are trying to detect. Below we show how this can be achieved by 'Winsorizing' a proportion of the top and bottom values. We then explore how overdispersion might be either multiplicative or additive: the additive model appears usually to fit data better and so is generally recommended. The significance of observed deviations then takes into account both the precision with which the indicator is measured within each trust and the estimated between-trust variability.

3.4. Winsorizing Z-scores

Winsorizing consists of shrinking in the extreme unadjusted Z-scores to some selected percentile, by using the following method.

(a) Rank cases according to their unadjusted Z-scores.
(b) Identify z_q and z_{1-q}, the $100q\%$ most extreme top and bottom unadjusted Z-scores, where q might, for example, be 0.1.
(c) Set the lowest $100q\%$ of z-scores to z_q, and the highest $100q\%$ of Z-scores to z_{1-q}. These are the Winsorized statistics.

This retains the same number of Z-scores but discounts the influence of outliers: an alternative would be to 'trim' the highest and lowest Z-scores.

3.5. Estimation of overdispersion

Following the standard approach of generalized linear modelling (McCullagh and Nelder, 1989) we first introduce a multiplicative overdispersion factor ϕ that will inflate the null standard error to $s_0\sqrt{\phi}$, i.e. we assume that the reason for the overdispersion is due to underestimation of the within-trust sampling error. Suppose that we have a sample of I units that we shall initially assume to be all adhering to a standard based on 'average' performance t. An estimate of ϕ is

$$\hat{\phi} = \sum_I z_i^2 / I \tag{2}$$

where z_i is the unadjusted Z-score (1) using t as the target, although Winsorized Z-scores may be used in estimating ϕ. A standard test for heterogeneity is given by the statistic $I\hat{\phi}$, which has an approximate χ^2_I-distribution under the null hypothesis that all units only exhibit random variability around the same underlying performance t. Overdispersion might only be assumed if $\hat{\phi}$ is significantly greater than 1, although generally such pretesting is to be avoided.

The 'adjusted' Z-scores are then given by

$$z_{\mathrm{D}} = \frac{z}{\sqrt{\hat{\phi}}} = \frac{y-t}{s_0\sqrt{\hat{\phi}}}.$$

The estimate $\hat{\phi}$ might additionally be multiplied by a debiasing factor based on the fact that, if all the institutions are in control and the only variability is due to overdispersion, the variance of the Winsorized Z-scores will tend to be less than ϕ (Spiegelhalter, 2005d). However, we have found that using this factor leads to somewhat wide limits and hence it has not been generally adopted.

3.6. An additive random-effects model

As an alternative to the multiplicative overdispersion model that was described above, we now consider an additive model in which each trust has its own true underlying level t_i with $E[y_i]=t_i$ and $\text{var}(y_i)=s_i^2$, so that $z_i=(y_i-t_i)/s_i$. An on-standard trust t_i is assumed to have a distribution with

$$\begin{aligned} E[t_i] &= t_0, \\ \text{var}(t_i) &= \tau^2. \end{aligned} \tag{3}$$

In other words the overdispersion is due to unexplained factors producing unavoidable variability between trusts, and so the standard is represented by a distribution rather than a single point. A standard method-of-moments procedure (DerSimonian and Laird, 1986) can be used to provide an estimate

$$\hat{\tau}^2 = \frac{I\hat{\phi} - (I-1)}{\sum_i w_i - \sum_i w_i^2 / \sum_i w_i}$$

where $w_i = 1/s_i^2$, and $I\hat{\phi}$ is the test statistic for heterogeneity: if $I\hat{\phi} < I-1$, then $\hat{\tau}^2$ is set to 0 and complete homogeneity is assumed. Otherwise the adjusted Z-scores are given by

$$z_{\text{D}} = \frac{y - t_0}{\sqrt{(s_0^2 + \hat{\tau}^2)}}.$$

Strictly speaking the standard error s_0, which is calculated assuming that the standard is true, should be adapted to allow for the standard no longer being a single point.

When constructing a funnel plot for multiplicative overdispersion, the control limits are adjusted to be $t \pm k s_0 \sqrt{\hat{\phi}}$, whereas for additive overdispersion they are $t \pm k\sqrt{(s_0^2 + \hat{\tau}^2)}$. If Z-scores have been calculated after a transformation, then control limits are calculated on the transformed scale, and then the axes should be labelled on the natural scale.

Fig. 3 shows an example comparing the standardized mortality ratios for all adult emergency admissions in 168 acute trusts in the third quarter of 2006. The highlighted outlying observation is Mid Staffordshire NHS Trust—this lies outside the upper 99.8% limits only for the additive overdispersion model. Of equal interest are the trusts that are identified as having better than expected performance.

From cross-sectional data alone it is difficult to distinguish between additive and multiplicative overdispersion, although techniques have been suggested (Lee and Nelder, 2000). The distribution of the points in a funnel plot can provide an informal guide, and in practice it is common to find that the empirical 'funnel' fails to continue to narrow for larger institutions, suggesting either inadequate risk adjustment or some small systematic differences between trusts: in either case additive overdispersion may be reasonable.

For longitudinal data the effects of additive overdispersion become clearer, as trusts that have high results tend to remain high over time: Section 4.5 shows how we can incorporate both forms of overdispersion when monitoring multiple series.

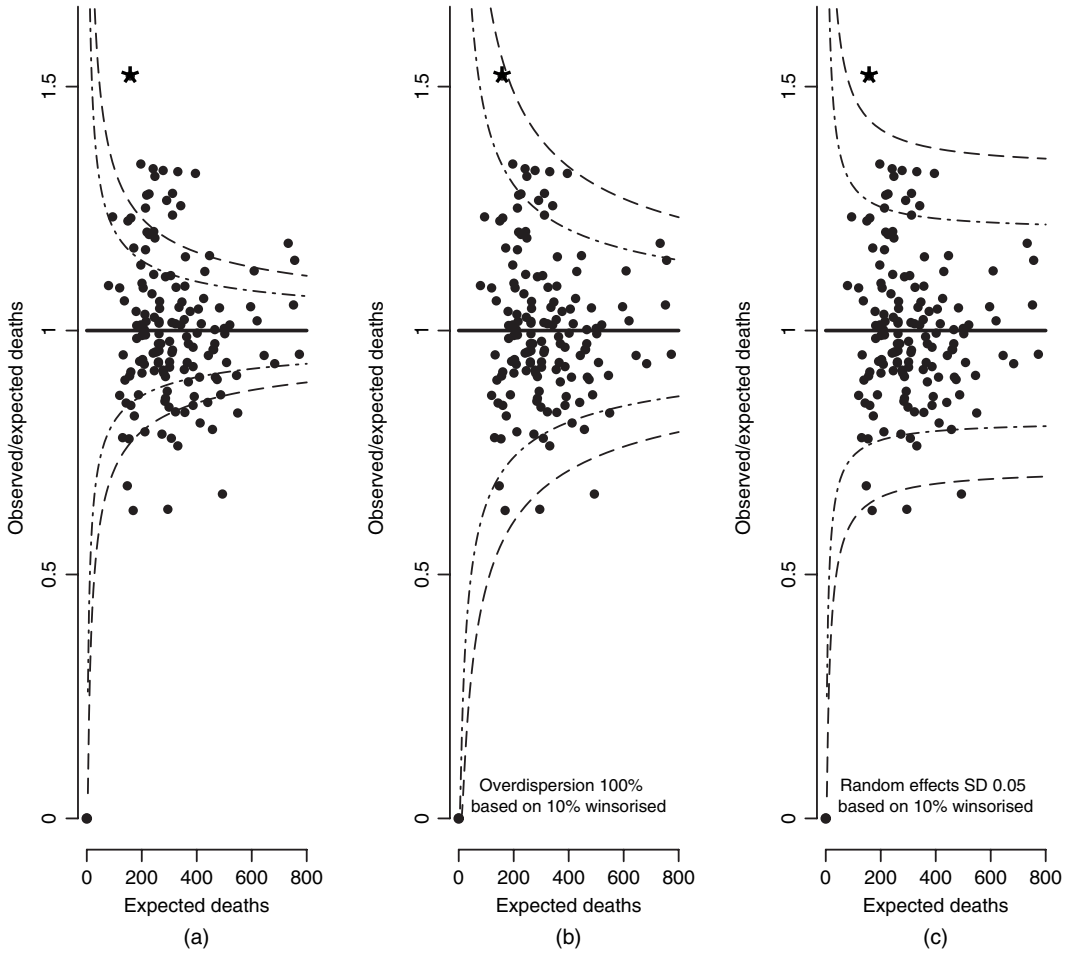

Fig. 3. Funnel plots with (a) no, (b) multiplicative and (c) additive overdispersion, showing the observed and expected deaths from all adult emergency admissions in the third quarter of 2006 (⁎, Mid Staffordshire Trust): overdispersion has been calculated following a square-root transformation (Appendix A.1) but plotted on the original scale; the percentage overdispersion is defined as $100(\sqrt{\hat{\phi}} - 1)$

3.7. Combining Z-scores

When combining the evidence from multiple related indicators, we want to aggregate a set of Z-scores $z_1, z_2, \ldots, z_n$, with (under the null hypothesis) mean 0, variance 1 and pairwise correlation c_{ij}. We may want to Winsorize the zs to prevent a single indicator having too strong an influence, e.g. by bringing all z-values beyond ± 3 back to ± 3.

Suppose that a column vector Z has expectation 0 and correlation matrix C. Then the most powerful test of a shift of the expectation to a vector *with all elements equal* is

$$D = (\mathbf{1}'TZ)/\sqrt{(\mathbf{1}'T\mathbf{1})},$$

where $\mathbf{1}$ is a column vector of 1s, and T is the inverse of the correlation matrix C: D has mean 0 and SD 1 under the null hypothesis and so is the appropriate composite Z-statistic. This can be expressed as

$$D = \sum_i \left(\sum_j t_{ij} \right) Z_i \Big/ \sqrt{\left(\sum_i \sum_j t_{ij} \right)}$$

where t_{ij} are the elements of T.

Although formally appropriate, this test would involve carrying out very large matrix inversions to obtain appropriate t_{ij}s, which is not feasible with the software that is available. The alternative is to use a different weighting, but one that discounts elements that are highly correlated with others. Now, for any arbitrary weights w_i applied to the Z_is, the aggregate Z-statistic with mean 0 and SD 1 under the null hypothesis is

$$Z_{\text{agg}} = \sum_i w_i Z_i \Big/ \sqrt{\left(\sum_i \sum_j w_i w_j c_{ij} \right)}$$

where c_{ij} are the correlations, with $c_{ii} = 1$. One option for Z_{agg} is the sample mean, leading to the composite Z-score

$$\sqrt{(n\bar{Z})} \Big/ \sqrt{\left(1 + 2 \sum_{i<j} \frac{c_{ij}}{n} \right)}.$$

This appropriately downweights sets of highly correlated Z-scores, while allowing low correlated Z-scores to 'reinforce each other': for example, if the Z-scores are uncorrelated the combined Z-score is $\bar{Z}\sqrt{n}$. However, experience has shown that this can lead to independent Z-scores having little influence, since essentially only the average correlation was being taken into account.

The current suggestion is to use as weights in Z_{agg}

$$w_i = 1 \Big/ \sum_j c_{ij}$$

where $0 \leqslant c_{ij} \leqslant 1$, i.e. if $c_{ij} < 0$ then it is set to 0.

3.8. Use of Z-scores in screening

As described in Section 3.1, items that were deemed relevant to a core standard were transformed to Z-scores, aggregated by using the techniques described above and cross-checked against each trust's self-declaration on that standard. The details of the exact process are complex but are summarized below—see Bardsley *et al.* (2009) for a fuller exposition.

Over 1700 specific data items from 35 source bodies were available in 2007–2008 to map to core standards: for example the safety domain comprised nine core standards (e.g. C01a—'Incident reporting') and for acute trusts 154 items from 40 different data streams were used to cross-check declarations in this domain.

For each trust and each standard, a label of 'potential undeclared non-compliance' was given based on a combination of the trust declaration, the Z-score and a subjective 'confidence score' based on an assessment of the indicator's quality, relevance and importance. Trusts were ordered according to the number of standards assessed as potential undeclared non-compliance and the top 10% of trusts were selected for inspection. A random sample of trusts that had not been selected for a risk-based inspection (originally 10% falling to 7.5% in 2007–2008) were also chosen for inspection, which provided a means to test the accuracy of the risk-based selection process. A certain level of random inspections also meant that there was the possibility of inspection even for those standards where there was little or no pre-existing information.

The inspection process was undertaken by local field staff using a distinct set of criteria and assessments, concluding with a judgement about whether the trust was 'compliant' with the standard. Standards for which a trust had declared compliance, but that were judged on inspection not to have been met, were categorized as 'qualified'. The success of the screening process was assessed by looking at the relative rates of qualifications in the risk-based and random inspections. In most, but not all, standards the qualification rate in the 'high risk' group was greater than in the 'random' group. Bardsley *et al.* (2009) showed that in 2006 the overall qualification rate in the high risk group was 26% and in the random group was 13% ($P < 0.008$, allowing for clustering). Where the random sample generated false negative results (i.e. the inspection found problems), the data used for that standard were used to inform the selection of items for the next screening round.

The screening process represented a systematic and objective approach to selecting trusts for inspection and was also successful in integrating information from many other national and regulatory bodies in an effort to maximize the use of data that had already been collected, while minimizing the burden of asking for new data. The importance of exploiting the rich variety of information within operational information systems is relevant to all regulators.

However, the system relies on cross-checking routine data against a public statement by the organizations themselves that they are or are not meeting required standards. This latter point has been criticized and seen as a form of self-assessment (Care Quality Commission, 2009c) though the aim was to make the declarations part of the trust's own governance procedures and to ensure that responsibility for compliance lies with the trust itself and not the regulator. In future the CQC will focus on registration of trusts, with *quality and risk profiles* bringing together multiple sources of data for 'estimating the risk of essential standards not being met' (Care Quality Commission, 2010). The statistical methods that were developed for screening are directly applicable to this function.

4. Surveillance

The screening that was described in the previous section was applied within an annual assessment framework, and a natural development is to use multiple heterogeneous sources of data for continuous monitoring of healthcare providers in the hope of detecting any problems as soon as possible. We term this process 'surveillance'.

4.1. Background

There has been substantial recent growth in the demand for surveillance of public health and clinical indicators. In the USA this has been primarily motivated by anxiety over bioterrorism following the anthrax attacks in 2001 (Bravata *et al.*, 2004) although any threats to human health may be included: for example, the Biosense Program of the US Center for Disease Control and Prevention (`http://www.cdc.gov/biosense/index.html`) is a major initiative in 'syndromic surveillance', where the aim is rapid identification of clustered outbreaks of disease. In contrast, in the UK the driving motivation behind surveillance of clinical care arises from 'scandals' such as the Bristol heart babies (Spiegelhalter *et al.*, 2002) and the Shipman murder case (Aylin *et al.*, 2003), and the public prominence given to hospital-acquired infections such as MRSA and *Clostridium difficile*.

The literature of statistical surveillance is huge and includes extensive coverage of the use of control charts within statistical process control: see for example Sonesson and Bock (2003) and Woodall (2006) for relevant reviews, and the papers in a special issue of *Statistics in Medicine*

(Fricker, 2011). Here we consider a range of procedures that are applicable between two extremes: from informal monitoring of a single series to a countrywide system featuring multiple indicators and hundreds of organizations, emphasizing reasonably simple procedures that nevertheless have recognized statistical properties.

4.2. Informal monitoring of a single series

Suppose that we observe over T time periods a series of repeated observations of an indicator within a single institution. These would typically comprise observed $O_1, \ldots, O_T$ and expected counts $E_1, \ldots, E_T$, e.g. MRSA *bacteraemia* counts and expected numbers in successive 6-month periods.

Informal monitoring requires presentation of the data in different formats without specifying formal thresholds that indicate when 'unusual' performance is detected. No single plot will be adequate and the following options could all be provided.

(a) *Longitudinal data summaries*: for example, superimposed plots of observed and expected counts can help communication and spot patterns.
(b) *Estimates of current performance*: this can be based on the raw data, or smoothed over the recent history by using, for example, an exponentially weighted moving average (EWMA) of the ratios O/E, or using a risk adjustment procedure (Grigg and Spiegelhalter, 2007).
(c) *Cumulative data summaries*: for observed and expected data, we can plot the cumulative observed minus expected events $(O - E)$ since a specific start time. This is also known as a variable life-adjusted display (Lovegrove *et al.*, 1999) or cumulative risk-adjusted mortality (Poloniecki *et al.*, 1998) plot. It is a useful display but is not formally optimal for detecting a step change (Grigg *et al.*, 2003), although limits from more efficient signalling methods can be superimposed (Sherlaw-Johnson, 2005).
(d) *Funnel plots of individual points*: these are useful to compare with other trusts but lose the ordering of the points unless they are connected in time sequence.

Fig. 4 shows an example for deaths following stroke in one of 147 trusts in each of 12 quarters between April 2005 and March 2008. The observed and expected numbers are shown as raw data, smoothed as an EWMA and superimposed on the EWMAs of the remaining trusts, the cumulative excess deaths, and a funnel plot in which each quarter provides a point and the observations for trust 62 (Mid Staffordshire NHS Trust) are shown. The funnel, without adjustment for overdispersion, fits the data very well.

The early performance for trust 62 is as expected, but there is a strong suggestion of an increase in the rate of mortality from quarter 10 (July–September 2007), with two individual points lying outside the 99.8% funnel.

4.3. Formal monitoring of a single series

Formal monitoring procedures for a series are based on a prespecified procedure for sounding an 'alarm', and hence require prior consideration of the basis for the alarm threshold to be used. There have been a number of recent developments concerning 'exact' methods for dealing with non-normal data such as Bernoulli, binomial and Poisson responses, e.g. risk-adjusted cumulative sums (CUSUMs) (Steiner *et al.*, 2000) and risk-adjusted EWMAs (Grigg and Spiegelhalter, 2007). Although elegant, the methods do rest on precise distributional assumptions and dealing with overdispersion, random effects, adjustments for multiplicity and so on can become technically complex. For robust routine use we have therefore found it appropriate to carry out preliminary transformations and then to use methods based on normal theory.

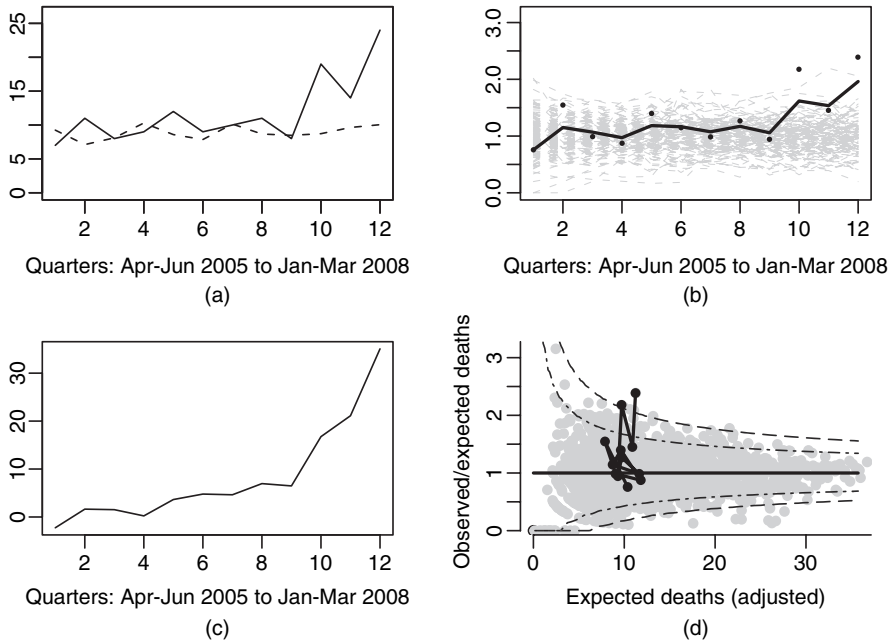

Fig. 4. Informal monitoring plots for observed and expected deaths from strokes, trust 62: (a) raw rates (———, observed; – – –, expected); (b) EWMA of O/E; (c) cumulative excess deaths (variable life-adjusted display); (d) funnel plot of O/E (– – –, 99.8% limits; · · · · · · ·, 95% limits)

Essentially, we monitor the 'naive' Z-scores (unadjusted for overdispersion) that were derived in Section 3 rather than the indicators themselves. These aim to provide approximate normality, and in particular to standardize variance within a trust across time. We note that the early aberration reporting system that was developed in Biosense also uses standardized Z-scores with expectations based on the very recent past (Tokars *et al.*, 2009): Szarka *et al.* (2011) investigate an alternative adaptive threshold method in which each observed statistic is converted to a P-value based on a historical empirical distribution for 'in-control' data, and then the P-value is converted to a Z-score by using the inverse standard normal cumulative distribution function.

To specify a monitoring system for each series, we need to specify the 'in-control' behaviour. For a particular indicator, assume that we have a set $z_t, t = 1, \ldots, T$, standardized measurements obtained by using the methods of Section 3, which may have been risk adjusted but which have not been adjusted for overdispersion since we are examining within-trust performance. We assume that, for an 'in-control' process, z_t has an approximate normal distribution

$$z_t \sim N(\theta, \sigma^2)$$

where θ is the target for the trust. Ideally we would like to assume $\theta = 0$ and $\sigma^2 = 1$, which would be so if the data were well behaved, but we allow for the possibility that the in-control within-trust variance σ^2 might not be 1, say reflecting multiplicative overdispersion due to clustering or insufficient risk adjustment. We also allow for the mean θ being non-zero, say reflecting additive overdispersion due to unavoidable systematic between-trust variability. With repeated data we shall have a chance to check these different sources of overdispersion, both within and between institution: Marshall *et al.* (2004) essentially assumed $\theta = 0$ and so monitored a multiplicatively overdispersed standardized measure.

Suppose for the moment that we have estimates or assumed values for parameters θ and σ^2 from which we can produce standardized Z-scores

$$z_t^* = (z_t - \theta)/\sigma: \tag{4}$$

these will have mean 0 and variance 1 under the null hypothesis of being in control. Formal testing can then be based on, for example, Shewhart charts that plot adjusted Z-scores $z_1^*, \ldots, z_T^*$ with thresholds set at, for example, ± 3, EWMAs of Z-scores (Szarka *et al.*, 2011) or the 'tabular' CUSUMs that are described below. Formal control limits could be set around the cumulative observed − expected (variable life-adjusted display) plot, but one must be wary of methods that 'build up credit' and lose sensitivity to recent changes in performance. Sherlaw-Johnson (2005) has shown how this can be avoided by mapping CUSUM limits onto the variable life-adjusted display, which is a technique that is widely used in Queensland, Australia (Clinical Practice Improvement Centre, 2008). An alternative control procedure for such plots (Sismanidis *et al.*, 2003) tests for recent changes but makes no allowance for the repeated testing taking place within each series.

4.4. 'Tabular' cumulative sums

Tabular CUSUMs plot the cumulative log-likelihood ratio where the likelihoods are conditional on a fully specified null and alternative hypothesis, constrained to lie above 0 (Steiner *et al.*, 2000; Grigg *et al.*, 2003). These are less intuitive than the informal methods but provide a more rigorous basis for concluding that a shift in performance has occurred. We shall consider the choice of an alternative hypothesis, the setting of thresholds, their interpretation and the transformation of an observed CUSUM to a P-value.

Following suitable transformation to a standardized Z^*-score (4), the null and alternative hypotheses are assumed to be

$$H_0 : z_t^* \sim N(0, 1),$$
$$H_1 : z_t^* \sim N(\delta, 1)$$

where δ is the alternative hypothesis that is discussed below. The log-likelihood ratio CUSUM contribution from the tth observation is then

$$\mathrm{LLR}_t = \max\{0, \mathrm{LLR}_{t-1} + \delta(z_t^* - \delta/2)\},$$

where $\mathrm{LLR}_0 = 0$. We note that the CUSUM never drops below 0 and only increases if $z_t^* > \delta/2$: i.e. the observed Z^*-score is at least half of the value in the alternative hypothesis.

A difficult issue is selection of an alternative. For fixed z_t^*, the maximum contribution to the log-likelihood ratio is obtained when $\delta = z_t^*$, i.e. an observation provides the maximum evidence for an alternative hypothesis set exactly equal to the observation. CUSUM charts are generally recommended for smaller δ that require cumulative evidence: Shewhart charts are sufficient for larger δ, say greater than 3. The current surveillance system takes $\delta = 2$, reflecting interest in performance more than 2 SDs above that expected.

Several criteria for setting both 'warning' and 'alarm' thresholds have been suggested: we could base them on the average run length before specific alternative hypotheses are detected, the power to detect a difference over a fixed run length, and so on—see Woodall (2006) and Fricker (2011) for discussion of metrics for comparing surveillance schemes. Grigg and Spiegelhalter (2008) have shown that the normal CUSUM can be transformed to a P-value under the null hypothesis of a steady state, and then standard methods used to decide a suitable threshold for the P-value. Default thresholds of 3 for alert and 5 for alarm have previously been suggested:

see Bottle and Aylin (2008) for a detailed discussion of use of risk-adjusted CUSUM methods for monitoring.

4.5. Handling multiple series

The majority of developments within statistical process control is concerned with monitoring a single series, with the recognized problem associated with repeated examination of accumulating data. In the healthcare context we have the major additional issue of multiplicity of institutions and indicators: for example the current surveillance system has around 1200 CUSUMs being simultaneously monitored in each of 160 acute trusts. This introduces severe problems in multiple testing and unexplained variability.

We consider a single indicator z_{it} being monitored in each of I trusts at time point t. We allow each trust i to have it own 'local baseline' θ_i, which we assume is distributed around a national standard θ_0, so that the whole model is

$$z_{it} \sim N(\theta_i, \sigma^2), \qquad \theta_i \sim N(\theta_0, \tau^2).$$

θ_0 should be 0 if using reasonable Z-scores, whereas τ^2 measures an 'acceptable' level of additive overdispersion. However, we note that this representation, in which the Z-scores are given a random-effects distribution, is not exactly compatible with the cross-sectional additive random-effects model in equation (3) which is based directly on the performance indicators. This shift to Z-scores is to try to ensure a reasonably common sampling variance for all trusts and all time points.

Grigg *et al.* (2009) described methods for estimating the parameters of this model from baseline data, say two observations on each trust. The variance parameters σ^2 and τ^2 should be estimated after Winsorizing, trimming or some other robust estimation method, to produce robustness against the very outliers that we are seeking to detect. Trusts could be stratified into types that might be expected to show similar variability over time, and a separate σ^2 could be estimated for each stratum. An empirical Bayes shrinkage estimate of the θ_is is appropriate to adjust for regression to the mean. Over time these estimates can be continually updated, whereby new parameters are based on a moving window of past data.

As Grigg *et al.* (2009) pointed out, two types of monitoring are then possible: a *local* comparison of each trust's data with its estimated local baseline $\hat{\theta}_i$, and a *relative* assessment of whether the current performance is divergent from the overall population of trusts. Local monitoring can be based on the standardization that is used for a single series in Section 4.2, so that

$$z^*_{it}(\text{local}) = (z_{it} - \hat{\theta}_i)/\hat{\sigma}. \tag{5}$$

Relative monitoring is more complex. One formulation of relative monitoring was provided by Grigg *et al.* (2009), who assumed the alternative hypothesis that trust i is not from the same population as the other trusts, but from a population that has been 'shifted' upwards by γ population SDs, so that

$$H_0 : z_{it} \sim N(\theta_0, \hat{\sigma}^2 + \hat{\tau}^2),$$
$$H_1 : z_{it} \sim N(\theta_0 + \gamma\hat{\tau}, \hat{\sigma}^2 + \hat{\tau}^2).$$

The contributions to the log-likelihood ratio are no longer independent under the null hypothesis, but Grigg *et al.* (2009) showed that a standard CUSUM based on

$$z^*_{it}(\text{relative}) = (z_{it} - \theta_0)/\sqrt{(\hat{\sigma}^2 + \hat{\tau}^2)} \tag{6}$$

can perform well: this is essentially the approach of Marshall *et al.* (2004).

Alternatively, we could test an 'extreme' simple null hypothesis for θ_i, say $\theta_H = \theta_0 + 2\hat{\tau}$. Then a relative Z-score

$$z^*_{it}(\text{extreme}) = (z_{it} - \theta_H)/\hat{\sigma} \tag{7}$$

could be analysed by using the standard CUSUM that was described above. This essentially tests whether a trust is operating at a level importantly above what is considered 'just tolerable'.

Fig. 5 shows the formal monitoring methods applied to the data that were informally plotted in Fig. 4. Alerts for both a change from baseline and systematic deviation from the population would be triggered in period 10, with alarms in period 12.

4.6. Thresholds for multiple series

Setting thresholds for many thousands of CUSUMs is a difficult task and many options have been tried by us and others. There are strong connections with areas in which vast numbers of hypothesis tests are being carried out, such as functional magnetic resonance imaging in brain scanning and microarray data for gene expression analysis, which have driven recent advances in statistical methods for dealing with extreme multiplicity. In all these areas we expect many null hypotheses to be false and so it is not appropriate to try to control the probability of a single false positive result occurring (known as the familywise error rate). Instead the aim is to control the proportion of false positive results out of the signals that are identified, which is known as the FDR. Jones *et al.* (2008) showed how the thresholds in funnel plots can be adjusted to control the FDR, whereas Marshall *et al.* (2004) looked at controlling the FDR over a limited time period.

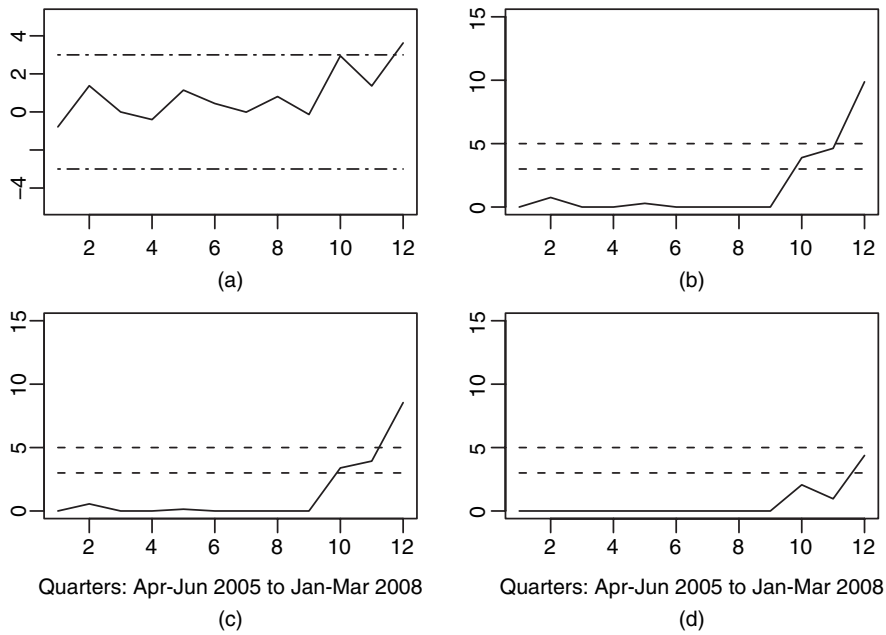

Fig. 5. Formal monitoring plots for observed and expected deaths from strokes between April 2005 and March 2008 previously shown in Fig. 4, trust 62 (CUSUM alert and alarm thresholds of 3 and 5 are shown): (a) Shewhart plot of the cross-sectional 'local' Z-scores (5) compared with the trust's own baseline; (b) 'local' CUSUM accumulating the local Z-scores against the trust's own baseline; (c) CUSUM *versus* 'population' accumulating the 'relative' Z-scores (6) compared with the overall population average; (d) CUSUM *versus* 'just tolerable' accumulating the 'extreme' Z-scores (7) compared with the high end of the population of trusts

Our preference is to turn everything into P-values (Grigg and Spiegelhalter, 2008) and then to use an FDR procedure as demonstrated in Grigg *et al.* (2009). However, in practice this means that the critical thresholds respond to the data and hence change over time, which can make the technique difficult to justify to a general audience. From experience we have found that using the P-value method to control the proportion of false positive results to below 10% generally arrived at thresholds near $p = 0.001$, and so this fixed choice has now been implemented.

4.7. Implementation issues for a surveillance system

Some of the potential difficulties in implementing a surveillance system in healthcare are revealed by considering the case of serial murderer Harold Shipman, a general practitioner who killed over 200 of his (mainly elderly) patients between 1975 and 1998, and was only finally caught because of a badly forged will. The resulting public inquiry chaired by Dame Janet Smith naturally considered whether he could have been caught earlier and commissioned a team to examine the statistical issues in detail (Aylin *et al.*, 2003), as well as taking additional evidence about formal monitoring schemes (Spiegelhalter *et al.*, 2003).

The statistical analysis showed clearly that Shipman could have been detected earlier if data had been collected and examined, in particular by using risk-adjusted CUSUMs allowing for multiple comparisons. However, a pilot investigation showed that, with only rudimentary risk adjustment, many doctors would signal with apparently raised mortality rates (Aylin *et al.*, 2003), which turned out to be due to blameless behaviour such as looking after elderly patients in care homes (Mohammed *et al.*, 2004). Nevertheless, Dame Janet Smith recommended that surveillance systems should be established saying 'in my view, the Department of Health must now make provision for it to be done' (Shipman Inquiry, 2004).

Clearly, one of the challenges of implementing a surveillance system is that, if not handled carefully, it could lead to false accusations generating unnecessary anxiety for patients and injustice to clinicians. Resolving these delicate issues is not simply a technical problem of choosing statistical methods and setting appropriate thresholds for action. Nevertheless, we do feel that it is vital that statistical methods play a strong part, in both

(a) the exploratory stage—alerting a human that something 'interesting' seems to be going on so that a planned follow-up strategy can be begun—
(b) the confirmatory stage—after careful data checking, to decide whether the data are sufficiently extreme to form part of the evidence in a formal review.

On the basis of these principles, in 2007 the Healthcare Commission began a process for generating and following up mortality alerts. This surveillance programme was reviewed after the first year (Care Quality Commission, 2009d) and a recommendation made that it be continued and expanded to other settings and a wider array of outcomes. This programme has been subsequently taken up by the CQC.

The surveillance system scans across groups of patients defined by their healthcare resource group on admission and triggering alerts based on a CUSUM approach. Given that there are about 1200 healthcare resource groups split between elective and emergency care and 160 acute trusts, there is potential for scanning around 200000 CUSUMs at a time. In reality, however, because of low numbers, some healthcare resource groups are grouped into baskets of related conditions. This process generates about 30 alerts every quarter. Other alerts for high mortality are received by the CQC from the Dr Foster Unit at Imperial College. These are generated by using a slightly different CUSUM method, where the risk of each individual patient is estimated from their clinical condition and characteristics on admission (Bottle and Aylin, 2008).

All alerts are then subject to internal scrutiny by the CQC to identify whether they are likely artefacts of the way that individual organizations record their data or due to casemix or organizational issues that are not adequately addressed by risk adjustment. After such an assessment, a decision is made whether or not to follow them up with the trust concerned. For example, there were 85 alerts between August 2007 and July 2008, of which 42 were pursued further with the trust (Care Quality Commission, 2009d). The engagement with the trust is initially in the form of a request for further information which in many cases can adequately explain the alert. On occasions, the alerts prompt trusts to review their own care processes, to recognize where improvements could be made and to respond with a plan of actions that hope to address these. Many of these are then monitored by CQC's regional compliance teams. In most cases alerts do not reappear after an engagement has taken place and sufficient time has been allowed for improvements to be implemented.

On rare occasions, a combination of multiple alerts and poor responses has led to an escalation of concerns. For example, Mid Staffordshire NHS Foundation Trust had been signalling on both Dr Foster's and the Healthcare Commission's surveillance systems since mid-2007 for a variety of outcomes: after repeated claims by the trust that these were due to coding difficulties an investigation began in March 2008 which led to conclusions of sustained excess mortality, particularly for emergency admissions, and a catalogue of organizational failures (Healthcare Commission, 2009).

The identification of a Mid Staffordshire Trust type of problem, however, needs to be balanced against the primary aim of surveillance as a system for identifying concerns within organizations before they reach such a stage. It is, perhaps, notable that in 2009–2010 approximately 30% of mortality alerts handled by the CQC led to an improvement plan being implemented by an NHS acute hospital. A list of all mortality alerts that have been closed as outlier cases, together with reasons for closure, are published every quarter by the CQC (Care Quality Commission, 2011).

5. Discussion

The NHS is one of the largest and most complex organizations in the world, employing over a million people. Over the past few years there has been an increased awareness and focus on the need for services that are effective and safe. Though both these concepts are difficult to measure unambiguously and it is impossible to ensure absolute safety or top quality in every process, it is reasonable to expect that systems should be in place that can rapidly spot where serious problems may be occurring.

The regulator's dilemma is then how to ensure the safety and effectiveness of services, yet without being overly prescriptive or intrusive to local staff. In addition a regulator needs to promote openness and accountability in healthcare such that assessments of the quality of care are accessible to the public—yet avoid making inappropriate and hasty judgements of complex issues. The annual health check had the problem that it was retrospective—a summary of what the regulator assessed the previous year or before. Reconciling these rather complex assessments with current experience in an organization is a major concern.

The surveillance process can, in contrast, be more topical. Formal monitoring clearly should have an important part to play in identifying emerging concerns, and in this paper we have suggested a range of techniques that may help in this complex task. The success of these methods depends crucially on the availability of complete, good quality and timely data. Ensuring such data can be complex and resource intensive, and unfortunately is not generally seen as a high priority. The burden on staff must be minimized and one way to do this is to exploit the

operational data that are collected within the systems themselves. In theory information that is useful for regulation should also be important for the management of local services.

A regulator needs to recognize that its own methods and data sets are open to scrutiny. The Healthcare Commission was proactive in publishing the details of the methods that it used in varying levels of detail, but there were situations where the timing of publication could potentially hamper the regulatory process and information was delayed. For example the screening data sets were not publicly released until inspections had been completed; rating scores were released for all trusts at the same time.

Of course the statistical issues form only one aspect of the challenge of improving the quality and safety of healthcare. There are major, and sometimes conflicting, demands from government, professionals and patients, and an increased expectation of transparency as demonstrated by the rise of Freedom of Information requests. This means that data collection and analysis, which have previously been viewed as rather a 'backroom' task to be carried out away from external scrutiny, are now at the forefront. Statisticians should relish this opportunity to show the value of their work.

Appendix A: Constructing *Z*-scores for different types of data

A.1. Indirectly standardized rates

Consider a standardized rate SMR = O/E based on an observed count O and expectation E. We assume that E is the target count, and a Poisson assumption implies that for the untransformed indicator SMR the standard is $t = 1$, with $s_0 = 1/\sqrt{E}$. Options for analysis include the following.

(a) Z-scores based on inverse normal transformation of exact P-values (Section 2.2): however, in poorly controlled circumstances the exact Poisson assumption is unlikely to hold for count data owing to overdispersion, and so reliance on these properties may be inappropriate.

(b) The untransformed unadjusted Z-score is

$$z = (O - E)/\sqrt{E}.$$

Suppose that we have two trusts which both have a standard of $E = 4$ adverse events, and one trust observed $O = 1$ and the other $O = 16$. They would receive Z-scores of -1.5 and 6 respectively, which strongly emphasizes the importance of the high compared with the low count.

(c) A logarithmic transformation to an indicator $y = \log(O/E)$ gives a standard $t = 0, s_0 \approx 1/\sqrt{E}$: this leaves open the question of what to do about 0 counts, and options include plugging in some small number, say minimum$(\frac{1}{2}, E/2)$. We have found that this transformation can overcorrect low counts, leading to a long negative tail that has an undue influence on critical limits for high counts when fitting a symmetric distribution. For example the trusts described above would have z-scores of $\pm\log(16) = \pm 2.77$ which gives the high and low counts equal weight.

(d) A square-root transformation so that $y = \sqrt{(O/E)}$ gives a standard $t = 1$, and $s_0 \approx 1/2\sqrt{E}$ and so

$$z = 2\sqrt{E}(y - 1) = 2(\sqrt{O} - \sqrt{E}).$$

Hence the trusts described previously would have $z = -2$ and $z = 4$, which seems reasonable. In theory the square-root transformation is approximately variance stabilizing if the sampling variance is proportional to the mean, which may not hold in general for count data, and some more complex transformations could be explored that might improve symmetry, normality and constant variance. However, we have found a simple square-root transformation robust to a wide range of empirical situations and has the benefit of being easily invertible.

A.2. Proportions

Consider an observed proportion r/n, with a standard proportion p. The binomial assumption implies $t = p$, and $s_0 = \sqrt{\{p(1-p)/n\}}$, and options again include

(a) Z-scores based on inverse normal transformation of exact P-values (Section 2.2),

(b) untransformed unadjusted Z-scores

$$z = \frac{r/n - p}{\sqrt{\{p(1-p)/n\}}}$$

and

(c) an empirical logit transformation.

The options have the same difficulties as outlined in Appendix A.1, and so we generally have adopted

(d) an inverse sine transformation $y = \sin^{-1}\{\sqrt{(r/n)}\}$, which gives a standard $t = \sin^{-1}(\sqrt{p})$, and $s_0 \approx 1/2\sqrt{n}$, and so

$$z = 2\sqrt{n}[\sin^{-1}\{\sqrt{(r/n)}\} - \sin^{-1}(\sqrt{p})].$$

A.3. Ordered categorical scores

Consider a categorical response which is ordered so that 'high is bad'. We adopt a normal scores approach by assuming that the categories are due to grouping of a 'latent' normally distributed quantity, so that the labels given to the categories are completely irrelevant, and all that matters is the proportion in each category. First the 'cut-offs' in a standard normal $N(0, 1)$ distribution are found that would give the observed proportions: for example, if a three-category response had observed proportions 70%, 20% and 10% in the three categories, this would represent cut-offs of 0.52 and 1.28, which divide a standard normal distribution into the required proportions. We then assign a Z-score to each category corresponding to the mean $N(0, 1)$ response within that category, using the result that, if $Z \sim N(0, 1)$, then

$$\exp(Z|a < Z < b) = -\frac{\{\phi(b) - \phi(a)\}}{\{\Phi(b) - \Phi(a)\}}$$

where $\phi(x)$ is the standard normal probability density, and $\Phi(x)$ is the standard normal distribution function. For the example above, the categories with proportions 70%, 20% and 10% are assigned Z-scores of -0.50, 0.86 and 1.75 respectively. We note that the final Z-scores do not depend in any way on the category labels or in regrouping of the categories.

A.4. 'Continuous' data

A minority of indicators are continuous, comprising mainly patient survey results and length-of-stay statistics. We assume an observed mean y with standard error s, with standard t. The unadjusted Z-score is

$$z = (y - t)/s;$$

if the data are provided in terms of a 95% confidence interval (LCL, UCL), then s may be obtained from $s = (\text{UCL} - \text{LCL})/4$.

A.5. Change data

Change is generally measured relatively and for indirectly standardized rates we shall have available an observed standardized rate ratio $\text{RR} = \text{SMR}_2/\text{SMR}_1$ with a target of t; for example t would be 0.8 if a 20% risk reduction were sought between period 1 and 2. If there is no overdispersion we can use the exact conditional procedure in Section 2.3, but in general we need to use a logarithmic transformation $y = \log(\text{RR})$, since any Z-score needs to be invariant to which indicator is used as the baseline denominator. The target is then $\log(t)$ and the standard error is taken as $s \approx \sqrt{(1/O_1 + 1/O_2)}$ (or, more accurately, we could have an estimate s_0 based on estimating the SMRs under the null hypothesis). If either O_1 or O_2 is 0 they can be changed, for example, to a suitable 'small' number such as min(non-zero O_1 or O_2)/10. We note that if O_1 and O_2 are of similar size then $s \approx 2/\sqrt{(O_1 + O_2)}$.

When considering changes in proportions we note that odds ratios are not generally used as a measure of change in this context and instead we consider a change from a proportion $p_1 = r_1/n_1$ to $p_2 = r_2/n_2$, with summary rate ratio $\text{RR} = p_2/p_1$. The indicator is $y = \log(\text{RR})$, and the standard error is taken as $s = \sqrt{\{(1 - p_1)/(n_1 p_1) + (1 - p_2)/(n_2 p_2)\}}$.

A.6. Ratios of counts

We may have available a ratio indicator of the form $r = O_1/O_2$ where O_1 and O_2 are counts, such as patients per general practitioner—this can just as easily be defined as general practitioners per patient. We

can adapt the methods that were described in Appendix A.5 for changes in SMRs by simply assuming that $E_1 = E_2$, although not with a target of 1. When expected counts E_1 and E_2 are available, we essentially have a ratio of SMRs, $r = (O_1/E_1)/(O_2/E_2)$.

A.7. Incorporating qualitative data

Intelligence derived from qualitative sources was also included in the data sets. These sources included commentaries from local stakeholders such as patient and public involvement forums, as well as information from investigations and the local operational staff of the regulator. Evidence from these sources was useful where it related to a specific core standard(s) and organization. In these cases evidence was subjectively coded onto a 'Z-like' score between –3 and 3, with a high positive score representing a strongly negative comment, using a simple coding framework describing key factors in the evidence. The coding results were subject to some test for interrater reliability. In some circumstances evidence could be given an overriding weighting. These cases were agreed by a small analytical panel.

References

American College of Surgeons (2010) American College of Surgeons Quality Improvement Programme. American College of Surgeons, Chicago. (Available from `http://acsnsqip.org/main/getstarted/documents/pdfs/ACS_NSQIP_2010_Brochure.pdf`.)

Aylin, P., Best, N., Bottle, A. and Marshall, E. C. (2003) Following Shipman: a pilot system for monitoring mortality rates in primary care. *Lancet*, **362**, 485–491.

Bardsley, M., Sherlaw-Johnson, C., Blunt, I. and Spiegelhalter, D. J. (2009) Using routine intelligence to target inspection of healthcare providers in England. *Qual. Safty Hlth Care*, **18**, 189–194.

Bird, S. M., Cox, D., Farewell, V. T., Goldstein, H., Holt, T. and Smith, P. C. (2005) Performance indicators: good, bad, and ugly. *J. R. Statist. Soc.* A, **168**, 1–27.

Bottle, A. and Aylin, P. (2008) Intelligent information: a national system for monitoring clinical performance. *Hlth Serv. Res.*, **43**, 10–31.

Bravata, D. M., McDonald, K. M., Smith, W. M., Rydzak, C., Szeto, H., Buckeridge, D. L., Haberland, C. and Owens, D. K. (2004) Systematic review: surveillance systems for early detection of bioterrorism-related diseases. *Ann. Intern. Med.*, **140**, 910–922.

Breslow, N. E. and Day, N. E. (1980) *Statistical Methods in Cancer Research*, vol. 1, *The Analysis of Case–Control Studies*. Lyon: International Agency for Research on Cancer.

Care Quality Commission (2009a) NHS performance ratings 2008/9. Care Quality Commission, London. (Available from `http://www.cqc.org.uk/_db/_documents/0809_NHS_ratings_overview_document_161009_200910164847.pdf`.)

Care Quality Commission (2009b) 2008/09 scoring methodology for existing commitments and national priorities. Care Quality Commission, London. (Available from `http://www.cqc.org.uk/_db/_documents/2008_09_scoring_methodology_for_existing_commitments_and_national_priorities_200910160652.pdf`.)

Care Quality Commission (2009c) Impact assessments on the costs and benefits to providers of regulated activities and people who use services: response to consultation on the impact assessment on the guidance about compliance. Care Quality Commission, London. (Available from `http://www.cqc.org.uk/_db/_documents/Response_to_consultation_on_the_impact_assessment_Dec_2009.pdf`.)

Care Quality Commission (2009d) Following up mortality 'outliers'—a review of the programme for taking action where data suggest there may be serious concerns about the safety of patients. Care Quality Commission, London. (Available from `http://www.cqc.org.uk/_db/_documents/Following_up_mortality_outliers_200906054425.pdf`.)

Care Quality Commission (2010) Quality and Risk Profiles of NHS trusts in early 2010. Care Quality Commission, London. (Available from `http://www.cqc.org.uk/_db/_documents/Quality_and_Risk_Profile_v0_FINAL.pdf`.)

Care Quality Commission (2011) Closed mortality outlier alerts. Care Quality Commission, London. (Available from `http://www.cqc.org.uk/aboutcqc/whatwedo/respondingtoconcerns/mortalityoutliers/closedmortalityoutlieralerts.cfm`.)

Clinical Practice Improvement Centre (2008) *VLADs for Dummies*. Milton: Wiley.

Department of Health (2004) National standards, local action: health and social care standards and planning framework 2005/6-2007/8. Department of Health, London. (Available from `http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4086057`.)

DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials. *Contr. Clin. Trials*, **7**, 177–188.

Financial Services Authority (2000) Our new approach to risk based regulation and what will be different for firms. Financial Services Authority, London. (Available from http://www.fsa.gov.uk/pubs/speeches/sp69_pres.pdf.)

Fricker, R.D. (2011) Some methodological issues in biosurveillance (with commentaries). *Statist. Med.*, **30**, 403–441.

Grigg, O., Farewell, V. F. and Spiegelhalter, D. J. (2003) Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Statist. Meth. Med. Res.*, **12**, 147–170.

Grigg, O. and Spiegelhalter, D. J. (2007) A simple risk-adjusted exponentially weighted moving average. *J. Am. Statist. Ass.*, **102**, 140–152.

Grigg, O. and Spiegelhalter, D. J. (2008) An empirical approximation to the null unbounded steady-state distribution of the CUSUM statistic. *Technometrics*, **50**, 501–511.

Grigg, O. A., Spiegelhalter, D. J. and Jones, H. E. (2009) Local and marginal control charts applied to methicillin-resistant *Staphylococcus aureus bacteraemia* reports in UK acute National Health Service trusts. *J. R. Statist. Soc.* A, **172**, 49–66.

Healthcare Commission (2007) Statistical banding for national target indicators, annual health check 2006/2007. Healthcare Commission, London. (Available from http://ratings2007.healthcarecommission.org.uk/Indicators_2007Nat/Downloads/Stats_banding_targets_methods_2_12-0.pdf.)

Healthcare Commission (2009) Investigation into Mid Staffordshire NHS Foundation Trust. Healthcare Commission, London. (Available from http://www.cqc.org.uk/_db/_documents/Investigation_into_Mid_Staffordshire_NHS_Foundation_Trust.pdf.)

HealthGrades (2010) Hospital Report Cards™ mortality and complication outcomes 2011 methodology. HealthGrades. (Available from http://www.healthgrades.com/business/img/HospitalReportCardsMortalityComplications2011.pdf#False.)

Health Protection Agency (2010) Results from the mandatory surveillance of MRSA bacteraemia. Health Protection Agency, London. (Available from http://www.hpa.org.uk/web/HPAweb&HPAwebStandard/HPAweb_C/1233906819629.)

Jones, H. E., Ohlssen, D. I. and Spiegelhalter, D. J. (2008) Use of the false discovery rate when comparing multiple health care providers. *J. Clin. Epidem.*, **61**, 232–240.

Kunadian, B., Dunning, J., Roberts, A. P., Morley, R. and de Belder, M. A. (2009) Funnel plots for comparing performance of PCI performing hospitals and cardiologists: demonstration of utility using the New York hospital mortality data. *Cathet. Cardvasc Intervn*, **73**, 589–594.

Lee, Y. and Nelder, J. A. (2000) Two ways of modelling overdispersion in non-normal data. *Appl. Statist.*, **49**, 591–598.

Lovegrove, J., Sherlaw-Johnson, C., Valencia, O., Treasure, T. and Gallivan, S. (1999) Monitoring the performance of cardiac surgeons. *J. Oper. Res. Soc.*, **50**, 684–689.

Marshall, C., Best, N., Bottle, A. and Aylin, P. (2004) Statistical issues in the prospective monitoring of health outcomes across multiple units. *J. R. Statist. Soc.* A, **167**, 541–559.

Mayer, E. K., Bottle, A., Rao, C., Darzi, A. W. and Athanasiou, T. (2009) Funnel plots and their emerging application in surgery. *Ann. Surg.*, **249**, 376–383.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Mohammed, M. A., Rathbone, A., Myers, P., Patel, D., Onions, H. and Stevens, A. (2004) An investigation into high mortality general practitioners flagged up via the Shipman Inquiry. *Br. Med. J.*, **328**, 1474–1477.

Poloniecki, J., Valencia, O. and Littlejohns, P. (1998) Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *Br. Med. J.*, **316**, 1697–1700.

Sherlaw-Johnson, C. (2005) A method for detecting runs of good and bad clinical outcomes on Variable Life-Adjusted Display (VLAD) charts. *Hlth Care Managmnt Sci.*, **8**, 61–65.

Shipman Inquiry (2004) Fifth report—safeguarding patients: lessons from the past—proposals for the future. Shipman Inquiry. (Available from http://www.the-shipman-inquiry.org.uk/5r_page.asp?id=4699.)

Sismanidis, C., Bland, M. and Poloniecki, J. (2003) Properties of the cumulative risk-adjusted mortality (CRAM) chart, including the number of deaths before a doubling of the death rate is detected. *Med. Decsn Makng*, **23**, 242–251.

Sonesson, C. and Bock, D. (2003) A review and discussion of prospective surveillance in public health. *J. R. Statist. Soc.* A, **166**, 5–21.

Spiegelhalter, D. J. (2005a) The mystery of the lost star: a statistical detective story. *Significance*, **2**, 150–153.

Spiegelhalter, D. J. (2005b) Funnel plots for institutional comparisons. *Statist. Med.*, **24**, 1185–1202.

Spiegelhalter, D. J. (2005c) Problems in assessing rates of infection with methicillin resistant Staphylococcus Aureus. *Br. Med. J.*, **331**, 1013–1015.

Spiegelhalter, D. J. (2005d) Handling over-dispersion of performance indicators. *Qual. Safty Hlth Care*, **14**, 347–351.

Spiegelhalter, D. J., Aylin, P., Best, N. G., Evans, S. J. W. and Murray, G. D. (2002) Commissioned analysis of surgical performance using routine data: lessons from the Bristol inquiry. *J. R. Statist. Soc.* A, **165**, 191–221.

Spiegelhalter, D. J., Grigg, O., Kinsman, R. and Treasure, T. (2003) Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *Int. J. Qual. Hlth Care*, **15**, 7–13.
Steiner, S. J., Cook, R. J., Farewell, V. T. and Treasure, T. (2000) Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, **1**, 441–452.
Straub, D. and Havbro-Faber, M. (2005) Risk based inspection planning for structural systems. *Struct. Safty*, **27**, 335–355.
Szarka, J. L., Gan, L. and Woodall, W. H. (2011) Comparison of the early aberration reporting system (EARS) W2 methods to an adaptive threshold method. *Statist. Med.*, **30**, 489–504.
Tenants Services Authority (2010) How we regulate: inspection. Tenants Services Authority, London. (Available from `http://www.tenantservicesauthority.org/server/show/nav.14478`.)
Tokars, J. I., Burkom, H., Xing, J., English, R., Bloom, S., Cox, K. and Pavlin, J. A. (2009) Enhancing time-series detection algorithms for automated biosurveillance. *Emergng Infect. Dis.*, **15**, 533–539.
Walker, A. S., Spiegelhalter, D. J., Crook, D. W., Wyllie, D., Morris, J. and Peto, T. E. A. (2008) Fairness of financial penalties to improve control of Clostridium difficile. *Br. Med. J.*, **337**, 1385–1387.
Woodall, W. H. (2006) The use of control charts in health-care and public-health surveillance. *J. Qual. Technol.*, **38**, 89–104.

Discussion on the paper by Spiegelhalter, Sherlaw-Johnson, Bardsley, Blunt, Wood and Grigg

Deborah Ashby (*Imperial College London*)

What are the indicators for a good discussion paper for a Royal Statistical Society Ordinary Meeting? Ideally it covers a 'big' issue for society, has statistically challenging aspects, generalizability to other areas and perhaps makes one reflect a little more deeply on current developments in statistics in its widest sense. This paper performs well on all indicators.

Is it a big issue? From time to time my father phones me to tell me about something he has read in the newspaper: Southend Hospital receiving three stars was one that sticks in my mind. Healthcare is important to people, so having some validation that local care is up to par matters. Equally, if there are problems, they need to be identified. So it is a big issue to him, which is good enough for me. But do the public trust these pronouncements? What do they actually mean? That is the thrust of this paper.

I said that I wanted some technical challenges. Many are addressed and I shall raise one more: that of assumed independence of units. Typically hospitals are ranked against each other, assuming that they contain sets of patients with no overlap. However, in some areas of medicine patients can be transferred between units, raising complex dependences between units, as well as issues of selection of patients into units. This is especially true in neonatal units, where babies can undergo multiple transfers. These units are quite rightly under scrutiny for performance, but the dependences need to be accounted for, and my colleague Shalini Santhakumaran has recently registered for a doctorate to explore this further.

How generalizable is this work? This paper is very much in the healthcare setting, but the authors allude to wider applicability. Have these ideas any parallels in the education sector in assessment of schools? Social services are an area of huge public interest, from child protection through to the increasing burden of care of the elderly. Could greater application of such principles help to maintain and improve standards without generating a huge bureaucracy? Do they have applicability in the financial sector to identify institutions at risk of failing?

I promised some wider reflections. My own work is largely in clinical trials and drug regulation—both the practice and the science of it. Not so long ago, it was shrouded in secrecy, and the usual cycle of events was that new drugs came onto the market surrounded by fanfare, and then sooner or later problems were reported, often leading to a 'knee-jerk' reaction to pull the drug from the market. There have been three big changes in recent years: the first is that we now work under much greater transparency, with assessment reports going into the public domain. The second is that doctors, other health professionals and patients can all report adverse events after taking medicines, and these reports are now routinely monitored by using principled statistical methods—every month the Commission on Human Medicines in the UK receives charts plotting reaction rates along with their empirical Bayes geometric means. The third is that there is an expectation that prospective monitoring is needed, and so drugs come in with a risk management plan, which means that further studies and appropriate surveillance are planned in. And a corollary, and possibly a cause, of all this is the much greater involvement of statisticians at all stages and levels in the process.

These changes have been hard enough in the regulation of medicines, which is a relatively well-defined area where statistical principles have for a long time received at least some recognition. What is interesting is how strongly they are paralleled by the developments that are charted in this paper, in territory which seems to me infinitely messier and more challenging.

The paper covers three different aspects: ratings based on measures that matter directly, such as mortality rates, screening for inspection that acts as an alert system to prompt a more intensive investigation and surveillance to pick up problems as they begin to emerge in order to do something about them and to correct them. I find a rather nice analogy here between terminal care of patients, treatment of disease and preventive measures. All three are necessary, but the shift towards prevention both in drug regulation and in healthcare regulation more widely is very welcome and made more possible by work such as that in this paper.

Professor Spiegelhalter and his colleagues have done a great service to statistics by engaging in these areas and then bringing together a scholarly exposition of these issues; the challenge is now to other members of the Society to engage equally in other domains, and to borrow from and to adapt the approaches in this paper. For doing this work, and bringing it so clearly to our attention, it gives me great pleasure to propose the vote of thanks.

Sheila M. Bird (*Medical Research Council Biostatistics Unit, Cambridge*)

This paper is a seminal reference on statistical methods in healthcare regulation. Many were developed by the authors, whose combined experience of their implementation at the Healthcare Commission (HC) is immense. Deep statistical thinking is translated for applications, and arguments virtuously veiled. Just getting 'consistent with target' adopted into practice (see Bird *et al.* (2005)) is no mean achievement. The Scottish Prison Service nearly had a riot—by governors—because of traducing this!

Other light touches leave a lesser legacy. I willed more on data quality, interpreting overdispersion, rational target setting, subject matter insights and learning cost effectively from inspections.

Rating organizations was by assessing performance against a *known* target. *Deciding cost efficiently whom to inspect*—not 'everyone every year' but risk based plus random (see Bird *et al.* (2005))—combined indicators, both quantitative and qualitative, to screen for potential undeclared non-compliance. *Quarterly* monitoring sought emergent problems *within* organizations.

But targets are not *known* if measured from an estimated baseline and their setting from on high may be awry as illustrated by annual methicillin-resistant *Staphylococcus aureus* set targets which reduced by 13 from baseline 63 to 24 over 3 years to give a 50% reduction!

Standardization was by readily reversed transformation to possibly overdispersed, Winsorized Z-scores that measure deviation from 'targets'. Overdispersion—multiplicative or additive—was typically disallowed when monitoring *within* organization. Overdispersion doubled the null standard error for observed *versus* expected deaths of adult emergency admissions whereas, for deaths after stroke, a covariate-adjusted funnel plot without overdispersion fitted well.

Other performance monitoring measures (see Bird *et al.* (2005)) are improve the quality of core data, reduce variation and garner subject matter insights.

Data quality: not all statisticians share the authors' relish for messy or noisy data. Some argue that time and treasure are better invested in sorting out the data—specialty by specialty—with subject matter and statistical expertise, deployed as a *joint enterprise*, so that data are 'trusted' sufficiently to yield subject matter insights. If not 'trusted' in that sense, are they sufficiently trustworthy to indict institutions or individuals? Costly, post-Shipman investigations into general practitioners whose practices with high mortality discovered only innocent explanations (mostly, a hospice for the dying) which, *a priori*, were surely far more likely than a *cadre* of general practitioners with murderous tendency.

The HC's final year cost £68 millions. By the standards of the National Institute for Health and Clinical Excellence that should buy 3500 quality-adjusted life-years through treatment or nearer 10 000 by prevention. Did it?

Risk-based inspections—being Bayesian: did feedback from inspections or screening alarms update the HC's prior belief that systematic bias (data quality or other reason) 'explained' outlying performance rather than 'excellence or direness'? Silence...: puzzling was the 26% *versus* 13% qualification rate in 10% risk-based *versus* similar number random inspections, as Bardsley *et al.* (2009) admitted that 5/44 standards to be checked per trust were chosen 'either by the highest risk estimates (risk-based) or at random (randomly selected visits)'. Did risk-based inspections yield a qualification rate that was higher than expected, or just 'higher than controls'? Was there a model-based expectation? If qualification rate were proportional to Poisson screening count with mean 5, and I assigned trusts with the highest 10% of screening counts to

risk-based inspection, their expected qualification rate would be twice that of trusts randomly selected from the remaining 90%.

Overdispersion—should reduce—but was accepted as a stratum-specific nuisance factor. When a specialty self-monitors, over time typically variation reduces *and* performance increases because best practice is shared and management improves through innovation. Did overdispersion reduce, and performance improve, over the HC's tenure? And, if so, was the improvement because of the HC or just monitored' by the HC?

Overdispersion—what is behind it: did the HC develop any sense of how much over dispersion was 'acceptable or expected' for different specialties? Tonight's authors could offer immense insight by analysing formally their ϕs or τs!

Quarterly alerts or alarms: unlike the HC's report on the Mid Staffordshire Trust (Bird, 2009), tonight's funnel plots properly display the performance of *all trusts* and dependence on whether overdispersion is multiplicative or additive. Designed so that a manageable 30 alarms 'pinged' per quarter, only 10% to be false discoveries, half the pings were resolved within the HC, apparently without notifying the trust. When notified, 30% of trusts put forward a remedial plan which took effect quickly, as mainly different indicators signalled in successive quarters. I commend a documentary account of HC quarterly alerts: which trust or indicator pinged, trust referral, how actioned—was data quality to blame?

False negative results: what has inspection delivered in a timely manner that was not forewarned locally? Shipman, Bristol, Mid Staffordshire Trust were late, forewarned locally, or both. The HC's investigations at St Georges and Papworth were timely because transplant teams chose to call them in. In a truly shared enterprise, trusts should inform the regulator about local issues that regulatory scrutiny has overlooked so that scrutiny improves. Did the HC achieve shared enterprise status?

With pleasure, I second the vote of thanks on a paper of outlying high quality whose methods, if widely dispersed, will be for the betterment of performance monitoring.

The vote of thanks was passed by acclamation.

Ian Hunt (*Edinburgh*)

Spiegelhalter and his colleagues set out to 'pay particular attention to the problem of simultaneously monitoring over 200 000 indicators for excess mortality' (first page). Unfortunately there is no mention of the main inferential problem caused by 'massive multiplicity': the ability of frequentist tests to discover interesting cases ('power'), for a given error rate, typically plummets as the number of hypotheses increases. Addressing power was a central recommendation from the Working Party on Performance Monitoring in the Public Services (Bird *et al.* (2005), pages 7–8). Do the manifestly frequentist methods that Spiegelhalter and his colleagues recommend provide the best, or even good, power to detect excess mortality? We are told that extreme cases like the Mid Staffordshire NHS Foundation Trust show up (page 20). But can the methods identify moderate, developing or chronic cases?

For initial surveillance purposes ('the exploratory phase') a frequentist is likely to investigate the indicators with the most extreme 'Z-scores' first. Formal power may not be such a concern: in recent times 30% of investigations have led to improvement plans (page 20). But formal 'confirmatory phases'—management review processes, court room proceedings and so on—require a higher mark of evidence than raw Z-scores. Without power a frequentist statistician is uncomfortable. She must formally convey that there is little statistical difference between many indicators and cases (see David Bartholomew's comments about publicly communicating uncertainty in Goldstein and Spiegelhalter (1996), page 428).

Would a Bayesian approach, dealing directly with the probabilities of hypotheses, be preferable? Certain Bayesians have said that the problem with multiple hypotheses 'does not really exist' (Lindley (1997), page 572) or just 'doesn't come up' (Gelman, 2009). But to make formal inferences a Bayesian must invoke a probability distribution over the entire hypothesis space. This can be problematic and controversial, especially with many related hypotheses (even basic probabilistic structures of hospital mortality rates are disputable (Austin, 2009)). 'Science is choking on the multiplicity problem' (Berger (2011), page 4). So, I fear, is mortality performance monitoring. And I do not think Bayesianism offers a methodological panacea.

So how might the ability to discover interesting cases be enhanced? First, consolidate the indicators into logical or meaningful groups—thus reducing the number of hypotheses while still using all the data (Efron calls this 'enrichment analysis' (Efron (2008), page 18)). Secondly, remove indicators that are unlikely to detect serious problems. Thirdly, be realistic about the limits of formal statistical inference methods (Bayesian or frequentist) and the evidential role of statisticians: real discoveries are made by human experts, not by mathematical methods.

Robert Grant (*St George's University of London and Kingston University*)

Professor Spiegelhalter and his colleagues have provided a lucid and comprehensive review of work on this topic in recent years. There are two aspects which I would like to focus on: transformation to Z-scores and the cumulative funnel plot.

The non-linear transformations to achieve approximately normal Z-scores are well known but a problem arises when combining multiple indicators into a single test for divergence as suggested in the paper (or equivalently forming a composite indicator). Indicators measured on different scales, such as standardized mortality ratios, waiting times and proportions, will have different non-linear transformations. The weights that are accorded to improvements on the various indicators are not fixed but rather functions of the observed level of compliance. A logistic transformation provides further reward for an improvement in hospitals already doing well (and, paradoxically, those doing badly) and little reward for the same improvement in those near the 50% mark. This could be seen as unfair or opaque, providing a clinician who is reluctant to admit mediocre performance with the excuse that they need to resist change, and clinical credibility is essential if our efforts are to bear fruit. The authors suggest that the fairest solution may be to weight indicators (or their Z-scores) according to the unique contribution that each makes to some metric such as generalized variance—essentially a principal components analysis—but this could also lack construct validity for clinicians as the weights could differ substantially from the perceived clinical importance. Exploring the effect of different weights on hospital results will encourage transparency and user engagement. A Monte Carlo simulation of weights ranging between the clinical importance, equal weights and those produced by a principal components analysis will easily provide a measure of uncertainty.

Secondly, I would recommend wider adoption of cumulative funnel plots which allow for the earliest possible detection of divergence from a standard of care and investigation of the reasons, in a readily understood format that explicitly displays the uncertainty in the estimates. The confusing path that is traced out in Fig. 4(d) would be untangled into a line moving from left to right as cases accumulate, if the plot was of standardized mortality ratios on the vertical axis, against the number of relevant cases on the horizontal. Such a plot also helps to spread awareness of the message that large numbers of data may be needed to compare performance, and that periods of organizational change can distort the results.

Thomas King (*University of Southampton*)

Two roles for statistics are presented: accountability and surveillance. The former is by design quite explicit but is acknowledged to be focused on failure (even if this is sometimes termed 'non-compliance'). There are two categories: 'failing' and 'not failing', with the statistical question being how to discriminate between the two by using the available data. The concept of the failure, too many deaths, too many infections, or even underachievement against too many indicators, can be considered accepted.

The remark that indicators might allow determination of success by being at the opposite end of the scale from failure is flawed. There is some evidence that public opinion is negative about institutions scoring badly against performance indicators but is disinterested in the distinction between mediocre and outstanding scores (Boyne *et al.*, 2009). Public accountability should be an inclusive construct and Young (2000) made it clear that engagement with corresponding narratives is necessary. If the public does not identify success with the significant absence of failure, then statisticians should not presume to do so on their behalf. Thus a separate, possibly locally constructed, indicator of success may be more appropriate.

Surveillance is conceived as a collaborative idea with early intervention and the best use of information at its heart. However, this requires a fair standard of statistical literacy at all levels of clinical governance which is thought not to be the case within the National Health Service (Advisory Committee on Mathematics Education (2011), page 22). Furthermore, this may conflict with the accountability if the same information is deployed with the expectation that a non-expert board can use it to demand executive action. Moreover, for this to foster public accountability, the public would need to understand the nature of this accountability, which is rather doubtful (King, 2011).

Statisticians are not so much presented with an opportunity as a responsibility: statistics can easily become rhetorical devices (Young (2000), page 79; Simpson and Dorling (l999)). Success and failure in the public sector are political constructs and statisticians should not attempt to construct political instruments for which they have no mandate. Our responsibility is to facilitate public understanding of the assumptions that are used in such instruments, giving a clear representation of the uncertainty they contain. The present paper is only a small step in this direction, leaving questions about identifying the cause of overdispersion and how to pursue formal accountability in the face of uncertainty.

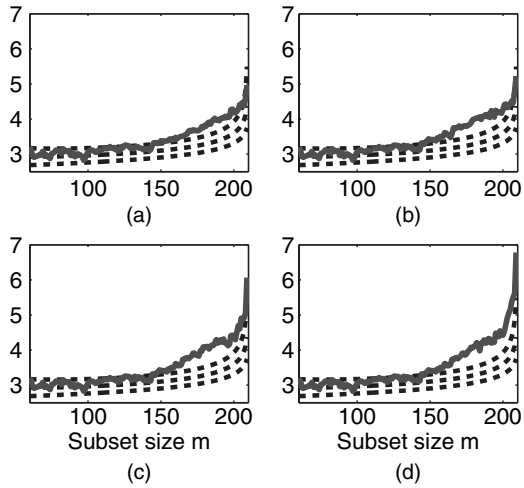

Fig. 6. Detection of outliers and unsuspected structure—forward plots of minimum squared Mahalanobis distances for inhomogeneous multivariate normal data of dimension 5; the samples contain 100 observations with variance 1, 100 with variance 3 and 10 with variance θ equal to (a) 3, (b) 5, (c) 7 and (d) 9; 1% and 99% pointwise envelopes; in (a) and (b) masking causes the largest value to lie within these bounds; for all values of θ departures are clearest around $m = 180$

A. C. Atkinson (*London School of Economics and Political Science*) **and M. Riani** (*Università di Parma*)
This interesting paper describes the work of many years on important and complicated problems involving much specific detail. We discuss the treatment of outliers (Section 3.4) and outline a second application involving monitoring and surveillance of large amounts of heterogeneous data.

The detection of outliers and unsuspected structure depends on obtaining parameter estimates free of departures from the null model. In the forward search (Atkinson and Riani, 2000; Atkinson *et al.*, 2004) subsets of the data of increasing size are used in fitting, starting from a subset that is robustly chosen to be outlier free; outliers enter towards the end of the search.

As an example related to the sums of squared Z-scores in the authors' equation (2), we simulated samples containing observations from three different five-dimensional multivariate normal distributions with independent observations: 100 with variance 1, 100 with variance 3 and 10 with the four different values shown in Fig. 6. As an outlier detection procedure, Riani *et al.* (2009) monitored the sequence of minimum squared Mahalanobis distances among observations that were not used in the subset for parameter estimation.

The plots include pointwise 1% and 99% bounds calculated from order statistics. Figs 6(a) and 6(b) show the results when the third variance is 3 and 5. At the end of the search, i.e. when all observations are used, the largest Mahalanobis distance lies within the bound, so, owing to masking, no outliers would be detected. However, the curve for smaller subset sizes shows systematic evidence of departure. This feature is similar in all plots; only in Figs 6(c) and 6(d), when the third variance equals 7 and 9, is the largest value outlying.

Masking is more severe in the analysis of mixtures of regression lines that occurs in investigations into the detection of fraud in international trade (Riani *et al.*, 2008). The data are simple regressions of price against quantity for a single good, but prices vary between firms; false declarations of price are used in tax evasion and money laundering. The challenging statistical problem is to disaggregate the data into regressions for each supplier, perhaps in the presence of outliers, and then to determine which transactions are fraudulent. As in monitoring the National Health Service, there is a vast amount of data which need to be analysed in a semi-automatic way. In addition, it is important not to create false positive results, since the mechanism of prosecution for fraud is cumbersome. If statistical methods are to be helpful they need to highlight cases from which successful actions can flow.

Axel Gandy (*Imperial College London*) **and Jan Terje Kvaløy** (*University of Stavanger*)
We congratulate the authors on this excellent paper which points to numerous further challenges. We

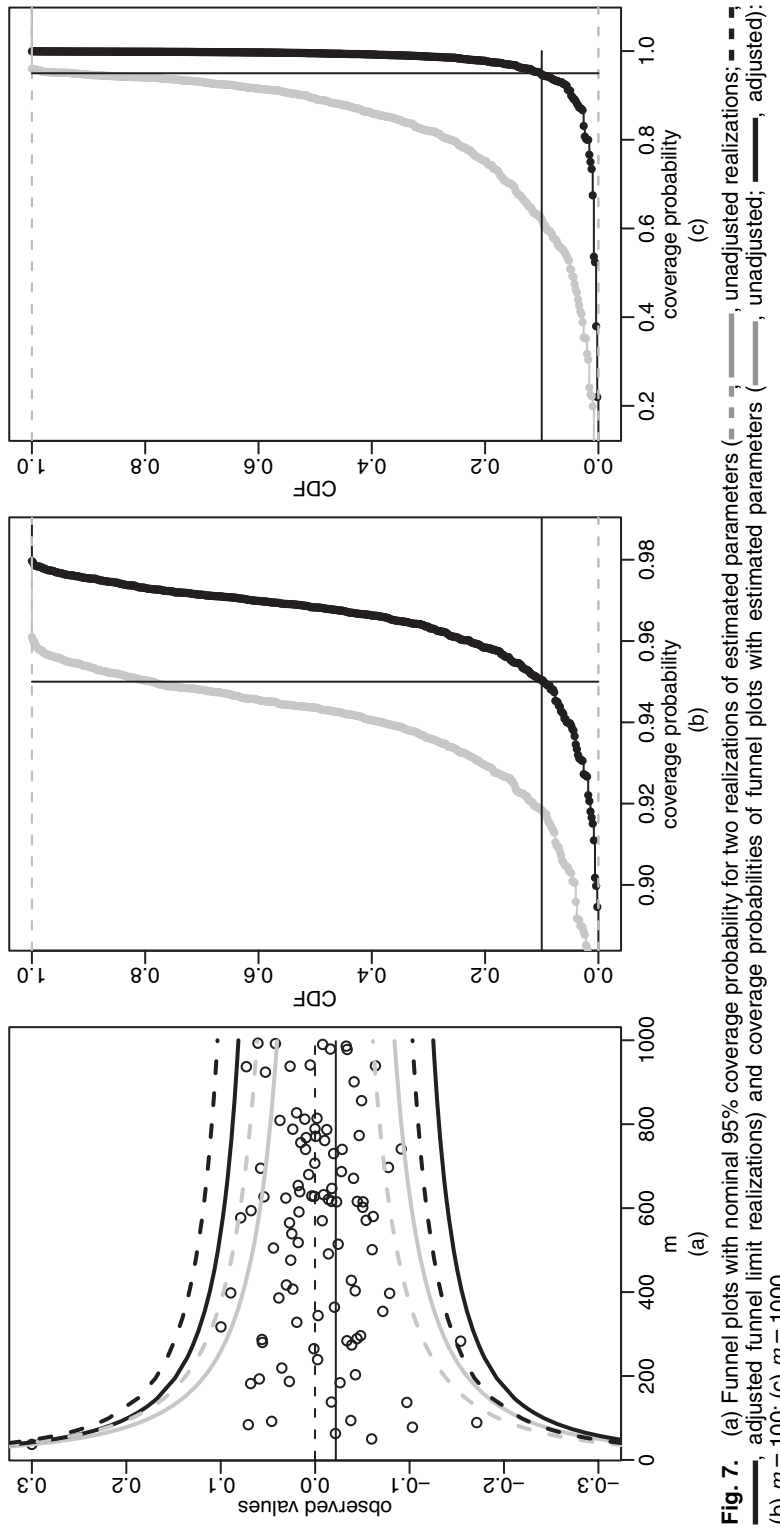

Fig. 7. (a) Funnel plots with nominal 95% coverage probability for two realizations of estimated parameters (– – –, ——, unadjusted realizations; ▬ ▬, ▬▬, adjusted funnel limit realizations) and coverage probabilities of funnel plots with estimated parameters (——, unadjusted; ——, adjusted): (b) $m = 100$; (c) $m = 1000$

comment on one challenge that was not discussed in detail in the paper: handling of estimation error of the target, or standard or in-control distribution.

As an example we consider funnel plots in the following simple situation. Suppose that observations are normally distributed and that we have n independent past observations $Y_1, \ldots, Y_n$ to estimate the unknown mean μ and variance σ^2 by the sample mean and variance $(\hat{\mu}, \hat{\sigma}^2)$.

For a hospital with m observations, the limits of the funnel plot based on the estimated observations are $\hat{\mu} \pm \hat{\sigma} k/\sqrt{m}$, where k is some quantile from the normal distribution. The actual coverage probability conditional on these estimates can be computed explicitly. It depends on $(\hat{\mu}, \hat{\sigma}^2)$ and is thus a random variable.

In Fig. 7(a) the grey curves are funnel limits based on two separate realizations of the past observations $Y_1, \ldots, Y_n$ with $n = 1000$, using $\mu = 0$ and $\sigma = 1$. The points are 100 realizations of new data with $\mu = 0, \sigma = 1$ and $m \sim \text{Unif}(\{10, \ldots, 999\})$. The horizontal lines are the estimators $\hat{\mu}$. For the broken curves, where the estimate $\hat{\mu}$ is close to the truth, most points lie within the funnel. For the full line, where the estimate $\hat{\mu}$ is a little off, more points than expected lie outside (mostly above) the funnel, i.e. we obtain too many false alarms.

We ran repeated simulations with the same set-up. The grey curves in Figs 7(b) and 7(c) are the cumulative distribution functions of the coverage probability for units of size $m = 100$ and $m = 1000$ respectively. Estimation has a considerable effect, and the impact of the estimation is worse for the larger m.

To correct the coverage probabilities of the funnels, one can construct pointwise tolerance intervals, e.g. guaranteeing that the nominal coverage probability is achieved in 90% of the cases. These tolerance intervals can be constructed via a bootstrap method similar to that described in Gandy and Kvaløy (2011) or by approximate methods. The black curves in all diagrams of Fig. 7 show the corresponding information for these corrected funnels. In particular Figs 7(b) and 7(c) show that the nominal coverage probability is achieved in 90% of the cases.

Obviously the influence of estimation will vary in different scenarios, but it will be an issue in all methods discussed in the paper.

Woody Caan (*Anglia Ruskin University, Cambridge*)

It is very useful to see in the paper how public bodies like the Care Quality Commission address health service needs like the 'rapid detection of emerging problems'. As a member of the Statistics User Forum, I wonder how official 'rating' will change when the government transparency agenda means publishing more data revealing the performance and quality of public services (Williams, 2011). Above all, will the behaviour of service users change (e.g. more litigation) or will the complexity of the regulation *processes* go right over the heads of the UK population, whose protection from harm is a regulator's duty?

This fascinating paper evokes two rather naive thoughts, in the mind of a former National Health Service manager who struggled with many a 'rather volatile indicator'.

The first is technical: the use of distributions such as Poisson. I collected many 'counts of adverse events' such as violence in hospitals (Powell *et al.*, 1994). Often these counts were not *independent* 'events', because of the phenomenon of the 'nightmare shift'. Once things began to wrong on the ward, a concatenation of escalating events was common, often involving the same people during a short period in one place. If a combustible mix of participants were present and something ignited them (e.g. racist taunts or petty bureaucracies) then violence bred violence. Aggression was not the only event that was clustered in such awful shifts. Later we found medication errors could often proliferate during one shift in one unfortunate ward.

The second thought is game playing by National Health Service organizations, which are well aware of the measures that are important for a good rating. Hypothetically if, say, a provider needed three stars for approval of a huge private finance initiative, but internally they knew current waiting list measures risked reducing their stars, then delaying such bad news to after the decision time for approval could be a temptation. Multiple examples of 'adjustment' of National Health Service data were first reported by the National Audit Office (2001). In the next few years government policy favours 'diversification' of health-care providers. Within that competitive arena, what 'adjustments' might managers in failing hospitals try to get away with?

Margaret Eames (*Acorns Public Health Research Unit, Hatfield, and Imperial College London*)

How can statisticians contribute to the new role of the Care Quality Commission (which replaced the

Healthcare Commission in April 2009) with its wider remit, to include *regulation of social care besides health* in England?

I am a statistician. I was Head of Public Health Intelligence providing health statistics to improve public health across Hertfordshire and Bedfordshire for 7 years, championing careers in public health intelligence for the National Health Service (NHS), and then working for the Healthcare Commission—as a public health development manager. I was in the Healthcare Commission preparing for the transition to the Care Quality Commission in 2009. Shaping the nature of the data collected for monitoring effective and safe social care nationwide is a challenge, perhaps more than health.

Social care regulation, alongside health as part of the Health and Social Care Bill, needs more careful thought, design and integration with health statistics regarding common definitions, for safety in the future. The linkage of social care with health could enable more effective regulation to prevent the abuses we have seen recently in the national media (e.g. the 'Baby P' death in Haringey, and the recent Panorama programme on Winterbourne View home, near Bristol).

Statisticians have an important role in designing objective questions, information technology systems and both quantitative and qualitative measures which can keep managers more accountable for the social care and service they deliver.

The measures themselves should include the effective linkage with health information in a locality, when requested to inform an enquiry. For example, with Baby P, it should have been possible for the data held on Baby P's injuries, by two different hospital accident and emergency departments, in the vicinity of Haringey, to be linked by using his NHS number. This was not connected (until too late, after he died).

When abuse is suspected this linkage should be a local routine public health intelligence information technology facility, enabled confidentially and safely for the protection of a vulnerable person, when the relevant social work department or whistle blower makes this request. This public health intelligence linkage could have saved a life in the case of Baby P. It has not been a priority in information technology for 'connecting for health' to enable this to date, but this could save lives in the future.

Hospital identification numbers (unique to one hospital) to identify patients are not enough—it is essential that the NHS number, now well defined and actively used by general practitioners, is recorded in hospitals to enable this linkage. Even if the 'whole NHS Spine' project is incomplete, this local linkage is very possible. Each local public health intelligence department (working with the local authority) could enable linkage, when several accident and emergency departments are accessible.

We could be more intelligent in defining appropriate standards of care, and procedures to enable prevention of some of the abuses that we have seen lately.

The following contributions were received in writing after the meeting.

Elja Arjas (*University of Helsinki and National Institute for Health and Welfare, Helsinki*)
In Section 2.2 the paper has the following interesting sentence:

> 'We have found it helpful to argue that targets should concern the underlying risk that is faced by patients, and the actual number of cases is only an imperfect measure of that underlying risk'.

I wonder whether it would be possible to formulate this understanding explicitly in terms of a statistical model, by viewing the problems of rating, screening and surveillance from the perspective of state estimation, or filtering. When suitably tuned, such a model could then also reflect what is believed about the dynamics of the underlying risk processes, for example, in how rapidly the risk levels in a considered unit or trust could be expected to change. A natural way to express the information that is contained in the latent risk level variables in terms of observables would be by issuing predictive distributions of the outcome variables that are being monitored, for example, annually and always predicting 1 year ahead in time. I should think that by making use of the extensive data that already exist, and possibly aided by simulation experiments for studying the sensitivity of the method, it would be possible to calibrate it to a level where such predictions would be realistic. More elaborate versions of the model could involve covariate information, and possibly allow for multivariate responses. A particular asset of these methods would be the concrete interpretation of their results: they would express, in terms of probabilities, what is to be expected next year when assuming that there is no outside intervention.

Dankmar Böhning (*University of Southampton*)
I congratulate the authors on a very interesting paper on continuous screening of healthcare indicators. I would like to comment on an important aspect and contribution of the paper, namely the construction

of an adjusted Z-score which incorporates an overdispersion estimate $\hat{\tau}^2$ based on the additive DerSimonian–Laird (DL) model. The estimate $\hat{\tau}^2$ plays an important part in the construction of the adjusted Z-score. Hence it seems desirable to have a reliable estimate of $\hat{\tau}^2$ available. In the case that the underlying data are standardized mortality ratios $\mathrm{SMR}_i = O_i$ it might be valuable to consider some of the estimators that were discussed in Böhning *et al.* (2004) as alternative to the popular DL estimator of τ^2. Here a class of unbiased estimators for τ^2 is defined as $\Sigma_i\, \alpha_i W_i / \Sigma_i\, \alpha_i$, where $W_i = \{(O_i - E_i\mu)^2 - E_i\mu\}/E_i^2$ and α_i are non-negative and non-random. Note that $E(W_i) = \tau^2$. This class provides a flexible family of simple, non-iterative and unbiased estimators for τ^2 which remains well defined under sparsity or zero counts O_i. Here μ is the mean over $\theta_i = E(\mathrm{SMR}_i|\mathrm{trust}\ i)$, the conditional mean of the standardized mortality ratio in trust i. To be more precise, it is assumed that θ_i has a distribution with mean μ and variance τ^2. If internal indirect standardization is used $\mu = 1$ necessarily; for external indirect estimation it can be estimated as $\Sigma_i O_i / \Sigma_i E_i$. Various weights are considered in Böhning *et al.* (2004) including equal weights, $\alpha_i = E_i$ and $\alpha_i = E_i^2$. In a simulation study $\hat{\tau}_2^2 = \Sigma_i E_i W_i / \Sigma_i E_i$ compared well with other estimators including the DL estimator. If there is still interest in using the DL estimator it is important to use the right form of the variances which are inversely involved as the weights w_i used in the construction of the DL estimator. Note that $\mathrm{var}(\mathrm{SMR}_i|\mathrm{trust}\ i) = \theta_i / E_i$ which could be estimated as SMR_i / E_i. However, this form is critical and leads to a breakdown of the estimator when $O_i = 0$. Instead, as argued in Böhning *et al.* (2002), a population-averaged version of the variance needs to be used, leading to $E(\theta_i / E_i) = \mu / E_i$ which contributes largely to the stability of the estimator and, I believe, is also used in the paper.

Michael J. Campbell, Richard M. Jacques, James Fotheringham, Ravi Maheswaran and Jon Nicholl
(*University of Sheffield*)

Recently our group in Sheffield have been evaluating a new summary hospital mortality index SHMI on behalf of the Department of Health (Campbell *et al.*, 2011). This is related to the Dr Foster hospital standardized mortality ratio HSMR but based on deaths in hospital and within 30 days of discharge from hospital. It uses only age, sex, type of admission and Charlson comorbidity as variables to standardize by, as the inclusion of other variables does not add additional discrimination in the performance of trusts. We were interested to see the contrast in the multiplicative and additive methods of defining the funnel plots for overdispersion that were given in Fig. 3 of the paper. We believe, on empirical and theoretical grounds, that additive overdispersion is probably more appropriate. Fig. 8 shows the expected deaths and SHMI from all admissions to non-specialist trusts for 2006–2007, and also shows Mid Staffordshire Trust as an outlier, along with four others. We can see that the points for SHMI do not really become much closer as the expected deaths increase, which suggests an additive model. Also we might expect uncertainty due to inadequate risk adjustment to add on the log-scale, as we usually model the data as a log-linear model for a Poisson outcome.

In view of the weight that hospital managers will attach to whether their hospital is above or below the line it is important to be clear on the calculations. We note that in Fig. 3 a square-root transformation was used. Is this based on an empirical inspection of the data or simply because a square-root transformation stabilizes variances for counts? Also in Fig. 3 we assume that you set $q = 0.1$. Does that mean that 20% of the points have been Winsorized? Do you have any views on the relative merits of trimming or Winsorization? In practice these lines will be used simply to decide, along with other information, which hospitals merit further inspection. Since additive overdispersion results in parallel warning lines as the expected values increase, this essentially reduces to a ranking exercise, except that some of the smaller hospitals can get away with higher SHMIs.

J. E. Chacón and J. Montanero (*Universidad de Extremadura, Badajoz*)

We congratulate the authors for this comprehensive work where several methodologies are considered with the aim of evaluating the efficiency of healthcare organizations. Depending on the specific goal, these methodologies are organized into three categories, which at the same time share some common elements and challenges.

Within the first category the goal is to analyse a single quality indicator from which, under some conditions, an exact distribution can be obtained in the case that the standard is met. The second category comprises the joint analysis of several indicators, seeking for a criterion to decide the trusts that will be inspected. The third category is oriented to monitoring an indicator with the goal of early detection of deviations with respect to a standard behaviour.

The common tools that are used to deal with these three categories of problem include several kinds

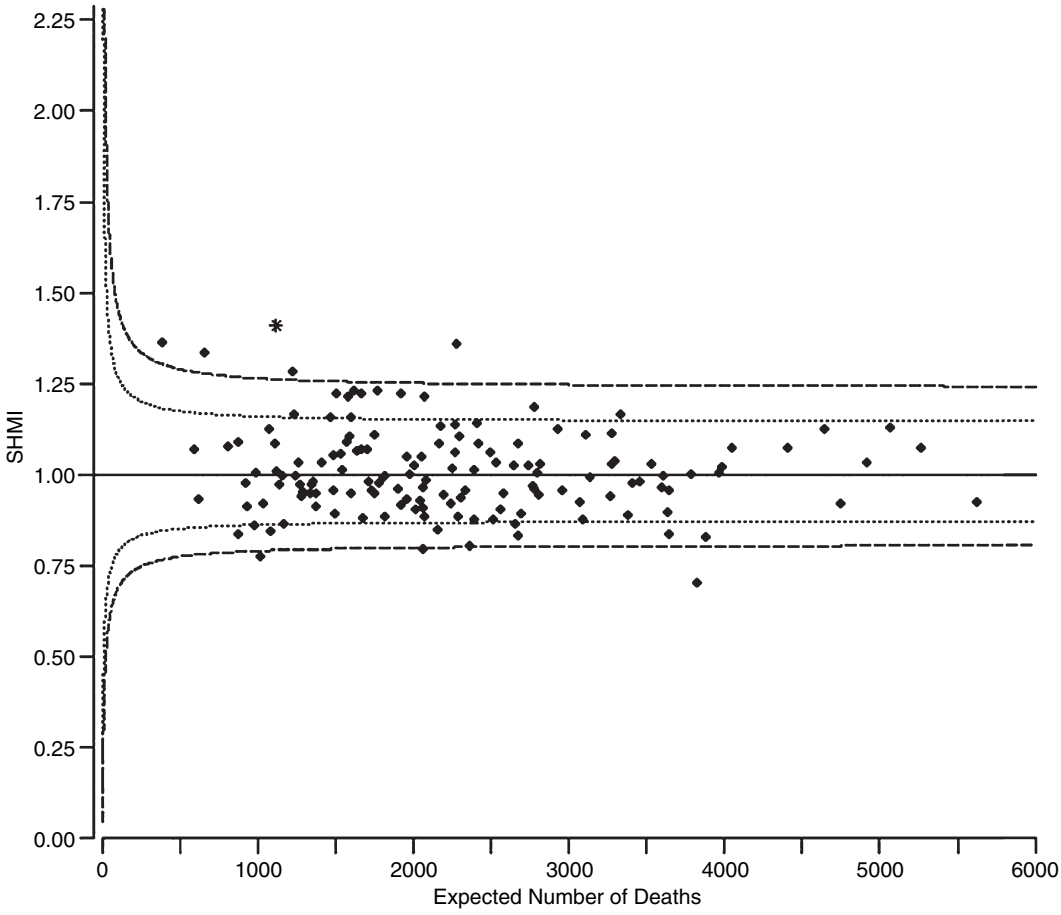

Fig. 8. Funnel plot showing the expected deaths and summary hospital mortality index from all admissions to non-specialist trusts for 2006–2007: ✳, Mid Staffordshire Trust

of Z-score adapted for overdispersion effects. These overdispersion effects occur when the observed variability is higher than would be expected, making most Z-scores appear extreme. Since 'expected' here refers to the underlying distribution, which is assumed to be normal, we wonder whether these discrepancies might be because the true underlying distribution could actually have tails that are heavier than normal.

In fact, as the authors themselves recognize (Section 4.3), 'some methods do rest on precise distributional assumptions', which involve mainly parametric models. We would like to draw attention to a variety of non-parametric approaches that may help to free us of such assumptions. For instance, funnel plots can be very useful for detecting outliers and abnormal behaviour, and Duval and Tweedie (2000a, b) introduced a non-parametric version through a trim-and-fill procedure which is based on rankings. Moreover, smoothing methods have also been shown to be useful to construct non-parametric analogues to classical Shewhart control charts. The properties of a non-parametric detection method, which was proposed by Devroye and Wise (1980), have been thoroughly examined in Baíllo *et al.* (2000) and compared with a parametric mixture model approach in Baíllo and Cuevas (2006). A further non-parametric alternative is based on the use of depth measures, as introduced in Liu and Singh (1993) and Liu (1995), possibly reducing a high dimensional multivariate problem to a univariate detection task on the basis of the corresponding data depth measures.

We feel that these non-parametric methods could provide a robust alternative in situations where clear deviations from normality are present.

Stephen E. Fienberg (*Carnegie Mellon University, Pittsburgh*)

As always, it is a great pleasure to read a paper that is focused on real public policy issues, where statisticians have been brought to the table and are making a difference. The authors identify several objectives for the use of outcomes data: ratings (and possibly rankings), screening and surveillance. In the USA, the Center for Medical Services has a related mandate to prepare outcomes measures for hospitals under the 2010 Affordable Health Care for America Act, for what we might characterize as a mix of ratings and screening, and some of the specifics of that mandate include establishing thresholds for outcomes to be used to reward or publish hospitals for their performance. Our methods also involve forms of standardized mortality ratios. Thus it was with great interest that I turned to this paper, especially to discover 'lessons learned' from the UK experiences.

Nonetheless, I was surprised in many ways by the heavy focus in the paper on computing Z-scores and P-values, and the relative absence of formal statistical models for risk adjustment except for that in the cumulative sum ideas. In much of the US research work on this topic, risk adjustment is carried out in the context of a formal statistical model, often based on individual level outcomes and their associated risks. Many of these models are hierarchical Bayesian models because of the natural way that such models capture different facets of the phenomenon of interest. Comparing US hospital outcomes across hospitals at the level of specific procedures leads to very sparse data arrays, far too sparse for the direct calculation of standardized mortality ratios without some form of smoothing. Bayesian hierarchical models provide one vehicle for such smoothing and the 'borrowing of strength' from similar facilities, however we choose to define them; for example, see Kipnis *et al.* (2010), Normand and Shahian (2007) and Rascz and Sedransk (2010). Given Professor Spiegelhalter's long-standing interests in and development of such models, their absence here is all the more surprising. Perhaps that is a consequence of how the results of the UK efforts are being used and by whom.

What are the principles that lead us to methods and models in this domain? Articulating principles and then developing formal methods that adhere to them but also work in practice is a great challenge. The present paper gives the statistical profession an opportunity to step back and to reflect on such principles as empirically driven efforts to regulate healthcare move forward.

Andrew Gelman (*Columbia University, New York*)

I applaud the authors' use of a mix of statistical methods to attack an important real world problem. Policy makers need results right away, and I admire the authors' ability and willingness to combine several modelling and significance testing ideas for the purposes of rating and surveillance.

That said, I am uncomfortable with the statistical ideas here, for three reasons. First, I feel that the methods proposed, centred as they are around data manipulation and corrections for uncertainty, have serious defects compared with a more model-based approach. My problem with methods based on P-values and Z-scores—however they happen to be adjusted—is that they draw discussion towards error rates, sequential analysis and other technical statistical concepts. In contrast, a model-based approach draws discussion towards the model and, from there, the process being modelled. I understand the appeal of P-value adjustment—many quantitatively trained people know about P-values—but I would much rather draw the statistics towards the data rather than the other way around. Once you have to bring out the funnel plot, this is to me a sign of (partial) failure, that you are talking about properties of a statistical summary rather than about the underlying process that generates the observed data.

My second difficulty is closely related: to me, the mapping seems tenuous from statistical significance to the ultimate healthcare and financial goals. I would prefer a more direct decision theoretic approach that focuses on practical significance.

That said, the authors of the paper under discussion are doing the work and I am not. I am sure that they have good reasons for using what I consider to be inferior methods, and I believe that one of the points of this discussion is to give them a chance to give this explanation.

Finally, I am glad that these methods result in 'ratings' rather than 'rankings'. As has been discussed by Louis (1984), Lockwood *et al.* (2002) and others, two huge problems arise when constructing ranks from noisy data. First, with unbalanced data (e.g. different sample sizes in different hospitals) there is no way to obtain reasonable point estimates of parameters and their rankings simultaneously. Second, ranks are notoriously noisy. Even with moderately large samples, estimated ranks are unstable and can be misleading, violating well-known principles of quality control by encouraging decision makers to chase noise rather than understanding and reducing variation (Deming, 2000). Thus, although I am unhappy with the components of the methods being used here, I like some aspects of the output.

Ronald B. Geskus (*Academic Medical Center, Amsterdam*)

As a medical statistician working in a different research area, I have read the description of the statistical methods in healthcare regulation with great interest. The issues and solutions are not only relevant if the unit under consideration is a trust, but also if it is a patient with a chronic disease. The development of personal health systems has made remote monitoring and treatment become increasingly important in chronic disease management. Information on the clinical condition of a patient is collected via stationary, portable or implantable devices, which can provide information on many variables of interest. For a single indicator (marker), it can be observed whether a threshold has been breached. Information from several variables may be combined to screen for patients with elevated risk profiles. Continuous monitoring allows for the detection of patients who show a sudden change in risk profile. For such patients, further inspection by a healthcare professional is needed.

The authors give a detailed description of the use and relevance of Z-scores. In the monitoring of performance against some standard, a trust can be observed to achieve, to underachieve or to fail. The calculation of these regions incorporates the role of chance. The authors explain the calculation of critical thresholds based on pointwise (e.g. yearly) methods. For example, in Fig. 1, the East Lancashire Hospitals Trust is observed to fail in 2006–2007. I wonder whether improvement can be obtained if the regions are determined on the basis of combined performance over the years. For example, if the number of methicillin-resistant *Staphylococcus aureus* infections had been consistently above the target, but still in the 'achieve' area in all three years, would that not be indicative of underachievement?

Unfortunately, later sections were more difficult to understand for someone who is not experienced in the methods that are described. Some statements are not explained in detail. It is not clear to me why a funnel that fails to narrow for larger institutions warrants the use of a model for additive rather than multiplicative overdispersion. In the section on surveillance, the cumulative sum method is given plenty of attention, but its rationale is not described in much detail. What is the difference between Fig. 5(c) and Fig. 5(d), apart from the choice of $\gamma = 2$ in the latter?

Hanna K. Jankowski (*York University*)

This work deals with the difficult topic of statistical methods in healthcare regulation, and I congratulate the authors on an interesting and thought-provoking paper.

One example given in the paper is that of performance monitoring for methicillin-resistant *Staphylococcus aureus bacteraemia* rates in trusts. With the desire of reducing the number of outbreaks, the objective was set of a 50% reduction in rates in 3 years, or a 20% reduction per year. The annual reduction was set as an absolute reduction relative to a single baseline rate. As mentioned by the authors, it is crucial that a robust baseline be established, and it is doubtful that the results of a single year would meet such a requirement for an individual trust. This issue was considered more extensively in a previous work of Spiegelhalter (2005).

A suggestion made by the authors is to consider instead an individual baseline by using data from a number of periods. As a simplification of the problem consider the following set-up: the number of cases in an individual trust is a Poisson process with constant rate λ in all previous years, $Y_{-1}, Y_{-2}, \ldots, Y_{-6}$. Under the null hypothesis that the trust has decreased their rate by 20%, the number of cases this year becomes Poisson distributed with rate $\lambda_0 = 0.8\lambda$. We consider the probability that

$$P(Y > y^* | \lambda_0) \tag{8}$$

and compare it with the same probability when λ_0 is estimated as 80% of the previous years' average, considering anywhere from 1 to 6 years into the past. The value y^* is taken as the critical value for $p^* = 0.841$. When the baseline rate is estimated on the basis of previous years, probability (8) was estimated based on $B = 100\,000$ samples. The results are shown in Fig. 9 for various values of λ. It seems that, in this simplified setting, at least 4 years are appropriate to reduce the additional variability caused by estimation of the baseline rate.

N. T. Longford (*SNTL and Universitat Pompeu Fabra, Barcelona*)

The paper reflects the view that statistics is about collating, quantifying and operating with evidence, i.e. with incomplete information. The evidence is used to assist the client (the Care Quality Commission) to decide which units to inspect, in what circumstances to apply contingency measures and whom to assign which grade. In an alternative perspective, the role of statistics is to make purposeful decisions in the presence of uncertainty (due to limited workloads of the units assessed). The qualifier 'purposeful' refers to serving the best interests of the client.

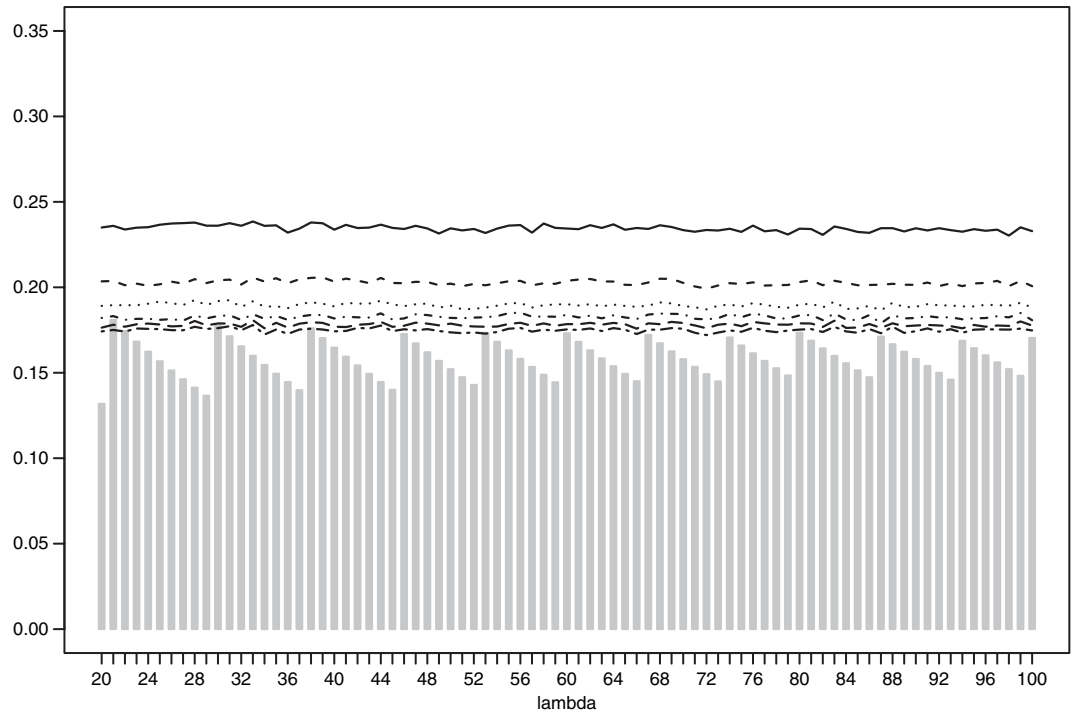

Fig. 9. True values (▩) of the probability (8) for various values of λ along with the probabilities when the baseline rate is estimated by using averages from previous years: the number of years in estimating the baseline is taken to be anywhere between 1 and 6 (———, 1 year; – – –, 2 years; · · · · · · ·, 3 years; · – · –, 4 years; — —, 5 years; · — ·, 6 years)

Central to an analysis that would be in accord with this perspective is a declaration of the losses due to the incorrect decisions that can be made. This imposes an additional burden on the analyst and the client, to elicit the loss functions (DeGroot, 1970; Lindley, 1985, 1998), or plausible sets of loss functions (Longford, 2010, 2011). With these functions, the solutions are actual proposed decisions, closely tailored to the remits and priorities of the client, avoiding arbitrary settings (the level of significance) and interventions in the analysis (being 'tolerant'). They ameliorate the issue of multiplicity, because (plausible) expected losses are evaluated instead of hypothetical probabilities. The non-statistical (qualitative) version of a declaration of loss functions is an integral part of *transparency* in how the Care Quality Commission conducts its business.

Within the confines of statistical methodology, it suffices to say that a *P*-value is singularly unsuited for making decisions (Lindley, 1998), because its derivation is not informed by its subsequent use—pursuing a course of action (or adopting a model subsequently regarded as valid). Hypothesis testing and similar procedures, including information criteria, are oblivious to the consequences of the erroneous decisions (Longford, 2005).

With the (plausible) loss functions, a unified methodology can be developed for screening, rating and surveillance, because the analyses for these three kinds of activities differ only by the features of these functions: false positive results are associated with relatively small losses in screening and high losses in surveillance, and rating represents an intermediate case.

An important conclusion of Spiegelhalter *et al.* (2002) is that the selection process, in this case the within-unit casemix distributions, may be a confounder in the comparisons of the units. It should have a prominent role in any analysis that aspires to compare like with like.

Thomas A. Louis (*Johns Hopkins Bloomberg School of Public Health, Baltimore*)

This paper is essential reading for all producers and consumers of healthcare regulation. I strongly endorse most of the authors' principal points and methods. I highlight a few of these and offer some elaborations.

Fair comparison: yes, a fair comparison 'must accommodate the vagaries of chance', but it also must use a valid basis for comparison and address the relevant target. Valid expected values are essential, and procedures must make an effective trade-off between signal and noise. When ranking, if uncertainties vary over units, hypothesis-test-based ranks tend to put the low variance units at the extremes; use of direct use of estimates has the opposite tendency. Best estimates directly address ranks, finding the appropriate balance (see Lin *et al.* (2006)).

Modelling goals: in developing expected values, the goal is not to find the 'best fit'; it is to produce values that support the counterfactual inference, 'if all hospitals treated the same mix of patients (or treated the same patients) how would they compare?'. Therefore, an appropriate expected value model accounts for patient-specific attributes at admission that associate with outcome; it should not adjust for post-admission events or hospital level attributes. These produce part of the residual variation and constitute the 'hospital effects' to be compared.

Appropriate targets: yes, '... targets should concern the underlying risk that is faced by patients, ...'. Consequently, statistical methods should directly address those targets. For example, estimates of ranks should be guided by a ranking-specific loss function; histogram estimates by histogram-specific loss. See Gelman and Price (1999), Lin *et al.* (2006) and Paddock and Louis (2011) for issues and examples.

Use of Z-scores: though use of Z-scores is prevalent, I encourage replacing them with estimates of target parameters (θs) with associated uncertainties (σ^2s). Happily, these can be produced from the Z-score and its denominator D: $\hat{\theta} = Z/D$; $\hat{\sigma}^2 = 1/D^2$.

Role of shrinkage: use of shrinkage estimates has been criticized for modifying the data (including changing ranks) and hiding outliers. It is challenging and essential that we demonstrate the benefits of careful hierarchical modelling and consequent shrinkage in producing inferences that are more stable and fair than those available from direct estimates.

Robustness: is absolutely necessary in collecting and preparing input data, in developing expected values and in developing comparisons. When comparing a large number of providers, robust Bayesian methods increase credibility of hierarchical modelling (see Lin *et al.* (2009) for an example).

Jorge Mateu (*University Jaume I, Castellón*)

The authors are to be congratulated on a valuable contribution and thought-provoking review paper. Healthcare regulation strategies set a key point in modern societies and receive important funding at national levels. The systematic collection of any kinds of data, and their corresponding analysis, has become essential to the planning and evaluation of public health practice. I shall focus my discussion on the context of space–time surveillance, which is a modern and timely topic in this area.

The authors focus basically on non-spatial, purely temporal methods derived from quality control ideas to monitor a stochastic process in time. In general control chart techniques are not sensitive to small changes in the process, although this is not so with the cumulative sum method (Frisén, 2003). However, these methods assume that data are independent in time, which is not a realistic assumption in many applications.

Realtime health outcome data are becoming more widely available and present interesting challenges for statisticians. Realtime epidemiology is the study of health outcomes in their natural setting and within a timescale, and it is often concerned with spatial (and/or temporal) variation (Lawson and Kleinman, 2005). Within this context, early and outbreak detection of anomalies in the observed pattern of incident cases is of key importance. Possible sources of data in a surveillance system come from general practitioners (with associated problems when reporting cases) and, in the UK system, from the National Health Service, which often gives date and location recorded for each call, but the spatial and temporal pattern of usage is unknown. In such cases, spatiotemporal statistical modelling improves the early disease outbreak and detection of anomalies (Diggle *et al.*, 2005). These statistical approaches must deal with the adjustment for temporal and spatial variation, the unknown time, place and size of an emergent cluster or the lack of suitable population-at-risk data (Assunção and Correa, 2009). As with screening and surveillance methodologies, classical methods for (space–time) cluster detection can be retrospective or prospective in nature. In the latter case, an events database is updated regularly and then an algorithm should be run to help to decide on the emergence of localized space–time clusters. In this latter approach, if one statistical test is carried out every time the database is updated, we face a severe multiple-testing problem with too many false alarms for clusters. Solutions can now be found in the literature (Diggle *et al.*, 2005; Assunção and Correa, 2009).

Kerrie Mengersen (*Queensland University of Technology, Brisbane*), **Tony Morton and Geoffrey Playford** (*Princess Alexandra Hospital, Brisbane*) **and Ian Smith** (*St Andrews Medical Institute, Brisbane*)

This paper charts a long journey of achievement in both biostatistics and healthcare. It clearly provides the foundation for further biostatistical innovation, such as more appropriate representation and interro-

gation of heterogeneity, improved methods for quantifying, combining and interpreting process measures and better ways to deal with multiple comparisons. However, we raise here a concern about the general direction of the journey itself, in particular the plethora of centrally mandated 'indicators' that have made these state of the art statistical methods necessary.

Our overarching aim is to make hospitals safer. This requires trust, which in turn requires transparency. Monitoring and reporting are necessary components of public trust. However, the public needs to know that the 'top-down' process of centrally counting adverse outcomes after they have occurred is indeed a cost-effective way of improving safety. Despite nearly 30 years of quality improvement implementation, there is much conflicting evidence (Millar, 2011; Houstein *et al.*, 2011; Runciman, 2011; Pratt, 2011; Jain *et al.*, 2011; Huskins *et al.*, 2011; Pronovost *et al.*, 2011; Benning *et al.*, 2011a, b; Landrigan *et al.*, 2010; Vincent, 2010). For years, the Princess Alexandra Hospital has released its annual infection report to the media: a 'bottom-up' initiative.

Trust and transparency are not only the mandate of the healthcare funders and consumers, but also of the healthcare practitioners. There are strong benefits in moving the monitoring process as close as possible (physically and temporally) to the point of care and carefully selecting relevant processes and outcomes (Smith *et al.*, 2011; Carthey *et al.*, 2001). As demonstrated at the Princess Alexandra Hospital and the St Andrews Medical Institute, if the motivation for this process comes from those delivering the care, and if it is internally constructed and managed, then unexpected changes can be detected relatively quickly; there is ownership; gaming becomes a non-issue; data error is minimized; data are used to learn how to do better rather than 'proving' compliance; people can concentrate on what they know makes a difference.

However, what shines through most is that good systems produce good hospitals. Modern hospitals are complex systems with emergent behaviour that depends on the interaction of a myriad of agents (Morton, 2011; Waterhouse *et al.*, 2011; Johnson, 2007). If we can better understand these systems, perhaps we can begin to influence this behaviour in more useful ways. To become safer, we have to take a positive approach and find out what causes 'safety'.

The paper thus brings us to a crossroads. We know that, to have safe hospitals, we must have safe hospital systems. Statistical analysis and modelling are fundamental to understanding these. However, instead of doggedly pursuing top-down indicator collection, the paper might equally motivate dedicated hospital surveillance staff and gifted statisticians to find new or complementary paths to creating safer hospitals.

A. F. Militino and M. D. Ugarte (*Universidad Pública de Navarra, Pamplona*)

This is a paper that deals with a real challenge for statisticians: to convince practitioners, health authorities, politicians and, possibly, the general audience of the necessity of using statistical procedures for healthcare regulation. In this sense, we find this paper very stimulating and we congratulate the authors for writing the paper to try to clarify some aspects about the role of statistics in this matter.

We see three important points here: first, how to define precisely the target, which clearly is not a statistical question; second, how to transform this target into numerical language and, third, how to estimate the indicators proposed. The authors mention anomalies in previous methods, but only because trial and error revealed discrepancies between evaluated indicators, and verified situations found in practice. But, how can we improve these indicators?: only by trial and error? Shall we be able to estimate bias, variances or any other dispersion measurements? In statistical terms only stochastic estimated error can provide assessment on the quality of the procedures. Lack of data makes it difficult to guess the goodness of these proposed indicators. Maybe some examples could clarify these items.

There are still a couple of points that we want to discuss. For example, the authors say that the precise methods discussed in the paper are not necessarily those that will be used in near future regulation duties. Is this particularly disappointing? Or, in contrast, is it more important to convince first on the necessity of simple and clear statistical analysis to be understood by non-experts rather than to offer 'universal' solutions to the regulation problems? As statisticians we would choose the latter argument. However, which would other people choose? Perhaps it is not so appropriate to change the procedures at the same time as the Commissions. The statistical procedures proposed should be simple, clear and as intuitive as possible. This could then contribute to their maintenance through time. We find that it is far more important first to educate people (the general public and politicians) about simple statistical concepts such as the notion of variability and uncertainty measures rather than offering 'sophisticated' statistical procedures.

Should we think about a supranational organization that monitors some chosen aspects of the European countries' national health services such as Eurostat or a similar organism: a type of European Health Organization? We find this option interesting for being able to compare national health services of the different European countries.

Which will be the instruments that allow us to compare some key performance indicators?

Greg Phillpotts (*Dorking*)
As commented earlier in the discussion, the aspect of data quality does, however, need further work. The paper covers overdispersion in Section 3.3, but this is only one aspect of the reality of data departing from theoretical good behaviour. In the National Health Service, information systems often depart drastically from the norms specified. Thus data for a subset of organizational units may not adhere to data collection protocols, possibly in relation to a subset of patients or to all their patients. This may result for example in what appears to be miscoding of case characteristics, or incorrect recording of reference dates or the wrong reference dates. Other data problems arise from the mutual inconsistency of apparently correct systems in different healthcare providers, or in subunits within providers. The source data that are required for comparisons of performance may have been the subject of many transfers between a wide variety of systems before being 'collected' for performance monitoring. Some of these problems result in haphazard errors in data, for which assumptions about random error may be appropriate, whereas many are systematic, resulting in bias, truncation or censoring, alone or in combination. These kinds of problem with data quality therefore in general give rise to problems *throughout* the distribution and not simply in relation to the apparent outliers that the work on overdispersion addresses.

As commented on by Professor Bird, the Society's Working Party report went further than providing criticism of performance indicators. The Working Party report also made some specific recommendations about performance monitoring regimes and it would be helpful if the authors could indicate how far this work takes forward that wider agenda.

In relation to surveillance, Section 4.7 comments that

> 'It is, perhaps, notable that ... approximately 30% of mortality alerts... led to an improvement plan being implemented'.

I suggest further discussion is needed before it can be seen as 'notable' that any particular fraction of alerts results in some specified action. There is to start with the chance of a false positive result, which may be high or low depending on the nature of the measure used and the system under surveillance. And, when an alert has been raised, the reality may be that only a small proportion of cases deserve a particular action plan. To give the full picture, examination is also needed of the specificity of the surveillance method as well as its sensitivity.

Emilio Porcu (*Universidad de Castilla la Mancha and University of Göttingen*)**, Carlos Alonso Malaver** (*Universidad Nacional de Colombia, Bogatá*) **and Alessandro Zini** (*Universitá Milano Bicocca, Milan*)
We congratulate the authors for this beautiful paper. We have some suggestions that may be considered by the authors for a more general view of the problem.

(a) The first issue is related to the use of P-values when the sample sizes are not sufficiently large and thus the asymptotic normality framework does not apply. De Martini (2008), among others, claimed that the usual classic testing procedures based on test variables (P-values, etc.) are less stable with respect to an equivalent approach based on the estimated power of the same test, the so-called reproducibility probability, for which constructive criticism can be found in Shao and Chao (2002), who noted that the reproducibility probability for a given clinical trial is useful in providing important information for regulatory agencies in deciding whether a single clinical trial is sufficient and for the pharmaceutical companies in adjusting the sample size in future clinical trials. Important evidence in favour of the reproducibility probability method can be also found in Goodman (1992), who complained that the P-value might lead to too optimistic interpretations of the test's result.
(b) We wonder whether seasonal or cyclical components should be kept in mind as a potential source of variability in the risk of contraction of illnesses.
(c) The assumption of normality is probably overly restrictive and it would be relevant to consider approaches that allow us to violate the assumption of normality, preserving a certain level of statistical efficiency.
(d) The score weighting scheme that is proposed by the authors is an important alternative that allows for a relevant computational gain (they avoid computation of the inverse of the covariance matrix) while preserving a good level of statistical efficiency. Although we advocate the use of such approaches and refer to a recent paper for them (Bevilacqua *et al.*, 2011), we consider that the scheme proposed is far too simplistic and restrictive. It would be important to consider a weighting scheme that allows us to take into account, for instance, geographical components, or the amount of a population or some other covariates. Some ideas may be picked up from Bevilacqua *et al.* (2011).

E. Marian Scott (*University of Glasgow*) **and J. Campbell Gemmell** (*Scottish Environment Protection Agency, Stirling*)

We congratulate the authors on a clear and stimulating paper. Although our focus is environmental regulation, there are many connections.

Regulatory framework

With respect to current environmental regulation, the regulator is a 'data collector and inspector', but new developments mirror 'a risk-based approach, being more targeted and proportionate'. Environmental regulation is largely about licensing and regulating economic activities that have the potential to pollute the environment. The 'norm' of environmental licensing today is based on four stages—assessment of the process and the risk that it presents, granting of permission to operate within clearly defined limits, monitoring and reporting on performance against the terms of the permit and enforcement. Modern regulation as described in the Hampton report (Hampton, 2005) advocates proportionate, risk-based permitting with lesser effort on site visits and monitoring in the light of risk assessment (and perhaps compliance performance history). This new regulatory practice places a *clear focus on results* (Sparrow, 2000) where indicators of success are measurable reductions. All of this resonates (although the language may differ) with healthcare regulation.

Screening and surveillance

When we consider surveillance and monitoring in the environmental context (with a variety of types of data, many hundreds of determinands and data quality issues), as well as monitoring of specific processes, monitoring is also undertaken to report on the state of the environment. In terms of their monitoring network, the Scottish Environment Protection Agency classify sites in three ways: *surveilliance*, *operational and investigative*. Of relevance here is *operational* monitoring, driven by risk assessments and located in areas of known risk and *investigative* monitoring which is responsive to unplanned events and emerging risks, where the source of the risk (the pressure) is not always well understood. As one example, the Water Framework Directive requires member states to classify water bodies dependent on ecological and chemical status. In Scotland this means classifying more than 3000 water bodies (or 25 000 km for rivers) annually (Scottish Environment Protection Agency, 2011).

Who is this for?

Stakeholders in the environment include governments, agencies, non-governmental organizations and the public. Environmental regulators must deal with considerable uncertainty, large volumes of data and complex environmental interconnectedness, but to communicate the state of our environment, and to license processes, our tools need also to be 'straightforward to implement, explanatory to multiple stakeholders and robust'.

The similarities, we argue, are striking.

Alfred Stein (*Twente University*)

I have read the paper with great interest, as it contains many novel elements that are to some degree well explained. At the moment I have the following issues.

The methodology is illustrated with examples from the British health system. It thus contains terms and notions that may not be available, or only available in a modified form, in other parts of the world. In the Netherlands, the numbers of trusts may be smaller than in the UK, and in the USA the number may be larger, whereas the health system in developing countries may be totally different. This also applies to the records. Although most of it may lead to only a minor change in the paper, I invite the authors to explain those concepts that are of a clear British character and hold these against the light of an international audience. This may include a brief explanation where modifications for use in other countries should be envisaged. Further, the paper does not address the spatial issue. I could imagine that in the British case a subdivision is made between the four countries, with four baselines instead of a single one. One may run into problems for Wales, for example, where I suspect that the number of trusts is substantially lower than in England, but it would be good to see the effects. The paper at this stage focuses on the number of deaths as an 'easy' measurement. This, without doubt, can be extended to prevalence. But I wonder whether also for a less-well-defined variable, i.e. for less 'easy' measurements, funnel plots are of use and provide important information. Possibly, the authors could comment on this.

In all, the paper is good to read and offers a range of possibilities for a further extension.

William H. Woodall (*Virginia Tech, Blacksburg*)

I congratulate the authors on developing a practical solution to a difficult problem. The false discovery rate approach to process monitoring, in particular, holds much promise in other applications. The US Centers for Disease Control and Prevention's 'Biosense' programme (`www.cdc.gov/biosense/`), for example, monitors such a large number of data streams that the excessive number of false alarms is a problem.

As far as technical comments are concerned, I only wish to point out some related work in the literature. Benjamini and Kling (1999, 2007) were the first to use P-values in statistical process monitoring, although their work was unfortunately never published. Lambert and Liu (2006) used a P-value-based approach in the monitoring of computer network data. On a different aspect of the methods, Hawkins (1993) used Winsorization to improve the robustness of cumulative sum charts to outliers in a way that is similar to that used by the authors.

The statistical approach that is used by the authors is a very good example of what Hoerl and Snee (2010) and Snee and Hoerl (2011) referred to as 'statistical engineering'. This concept has been receiving much emphasis in the USA over the last couple of years. At their outset statistical engineering projects are characterized by the following seven characteristics.

(a) The solution will satisfy a high level need.
(b) There is currently no satisfactory solution to the problem.
(c) The problem has a high degree of complexity involving both technical and non-technical challenges.
(d) More than one statistical method is required for solution.
(e) Long-term success requires embedding the solution into work processes through customized software.
(f) The influence is greater than could be achieved with individual tools.
(g) The solution can be leveraged to similar problems elsewhere.

Statistical engineering projects appear in many fields. Those promoting the concept of statistical engineering argue that study of these types of projects is needed to identify common success factors. Also, they believe that statistics students should be taught about these practical, high impact projects that require statisticians to go beyond what is typically presented in textbooks. The value of this framework is yet to be fully determined, but I believe that it is worthy of discussion within the statistical community.

The **authors** replied later, in writing, as follows.

We are very grateful for all the comments, whether complimentary or critical. We have tried to group our response into broad areas.

The basic statistical philosophy

Professor Fienberg raises two separate but vital issues in contrasting our approach with the work of the US Center for Medical Services. The first concerns risk adjustment. This is barely discussed in our paper but remains implicit in all our analyses of observed and expected events: the expected counts should be derived under whatever risk adjustment model is available.

The second issue, which is also raised by Professor Gelman and Professor Louis, concerns the benefits of a 'model-based approach' in contrast with our use of Z-scores and P-values, and in particular hierarchical Bayesian modelling in which centre-specific parameters are estimated by using shrinkage methods and thresholds are based directly on the estimates of these underlying parameters. We would claim that we do use a model-based approach, in that our additive random-effects models are essentially (empirical) Bayes hierarchical models. The major difference is that instead of estimation we use a hypothesis testing framework for identifying outliers. This is deliberate. In the estimation framework, the centre parameters are assumed to be exchangeable and drawn from the assumed random-effects distribution, and 'extreme' cases are identified by examining the posterior probability that a centre parameter exceeds some critical threshold. Note that it appears inappropriate to use as a measure of extremeness the posterior probability that the centre parameter is greater than the overall mean, as used, for example, by Rascz and Sedransk (2010). In an additive random-effects model this is simply the chance that the institution is worse than average—that it ranks in the worst half of the distribution, which is not of much interest given that we know that half of all institutions are worse than average. Shrinkage can also obscure outliers, as pointed out by Professor Louis.

Rather than using a shrinkage estimate based on an assumed encompassing model, we are really interested in whether the institution is 'divergent', i.e. it is not drawn from the overall distribution that describes

the majority of the institutions. This is a test of a distributional hypothesis and can be based on a *P*-value—this approach essentially corresponds to the Bayesian *P*-values that are promoted by Professor Gelman, although admittedly we do not fully allow for parameter uncertainty when creating these *P*-values. See Jones and Spiegelhalter (2011) for a full discussion of these issues.

Professor Gelman also argues that statistics should be drawn towards the data. We agree, which is why we much prefer funnel plots of the original data rather than portrayals of shrunk estimates with intervals, which are likely to be met by considerable suspicion. And, as Gelman and Louis emphasize, rankings are to be avoided.

Professor Longford and Professor Gelman suggest a full decision theoretic approach with losses for different decisions. Some implicit loss function lies behind the setting of thresholds and trade-off between different types of surveillance errors, but we feel that formalizing this would be a step too far for our collaborators. We acknowledge Ian Hunt's doubts that full Bayesian methods would really solve the multiplicity problem, and we agree that it is more appropriate to aggregate indicators further and acknowledge limitations of the method.

Specific methods used

The discussants have made numerous suggestions for improvements of the specific methods adopted, and we would readily admit that we have at each stage tried to select the simplest possible technique and refinements are undoubtedly possible. Robert Grant raises the issue of non-linear transformations and their influence on appropriate weighting. We only estimate weights for screening, and we feel that if weights are to be used as part of an accountable judgement then they should not be obtained statistically, as they are value judgements that need to be explicit and not hidden behind statistical niceties. But we agree that cumulative funnel plots are attractive although they make no explicit allowance for multiple testing.

Some improvements to estimating our 'in-control' null distributions are proposed. Professor Atkinson and Professor Riani recommend robust estimation of an in-control process by using a sequential inclusion of cases, and we agree that there are probably better ways than our simple Winsorization. Axel Gandy and Jan Terje Kvaløy point out that ignoring estimation error in setting a standard can lead to an anticonservative procedure, but we use plug-in values since in principle the standard should be a fixed quantity, externally set, rather than an estimate of some underlying parameter. Professor Böhning's improved estimators of the random-effects parameters appear admirable since they are straightforward to implement but, although the non-parametric methods for funnel plots that were suggested by Professor Chacon and Professor Montanero are a possible replacement for our transformation approach, we have worked under severe constraints of simplicity and feel that they would be too complex: similarly some of the excellent suggestions of Dr Porcu and his colleagues may be too sophisticated for the purpose in hand.

The work on the new SHMI of Professor Campbell and colleagues is very exciting and provides support for additive overdispersion. In answer to their query, we use a square-root transformation of counts on both theoretical grounds and because in practice it seems to produce reasonably symmetric 'funnel-shaped' distributions.

Professor Caan correctly identifies potential lack of independence of events arising from, say, outbreaks of methicillin-resistant *Staphylococcus aureus*, although allowance for overdispersion should provide adjustment for such clusters. Non-independence of units due, say, to transferred patients presents a greater problem, as Professor Ashby identifies.

Professor Arjas suggests making targets based on explicit predictions under a null model, with full allowance for uncertainties. This is certainly attractive, but we would still be faced with specifying bounds (e.g. *P*-values) for identifying 'extreme' and 'divergent' institutions. Finally, Hanna Jankowski's nice analysis suggests that historic baselines of 4 years may be appropriate for stable setting of methicillin-resistant *Staphylococcus aureus* type targets, whereas Ronald Geskus suggests that consistently just subnormal performance should be identified—this would not happen in the annual scoring system but is exactly what surveillance is supposed to pick up since cumulative sum methods accumulate such deviations.

Generalization

Our aim has been to develop methods that may be applicable to a far wider range of contexts than those in which we have been engaged. We are therefore delighted that connections with other areas have been drawn by the discussants. Professor Ashby suggests that social services could benefit, although Margaret Eames points out that lack of linkage across institutions may prevent individual cases from being identified—again an issue of non-independence. Alfred Stein asks about generalizations to other healthcare

systems, which clearly will differ with respect to their funding and regulatory responsibilities. However, the ideas seem generic—multiple institutions, variations in casemix, multiple outcomes and indicators, fairly simple statistical techniques and the need to identify divergent behaviour as soon as possible. Professor Mateu points out that space–time monitoring also faces the same requirements of robust estimation of in-control processes and allowance for multiplicity.

The strong connections that Professor Scott and Professor Gemmell make with environmental regulation are compelling and suggest, in a way that was completely unappreciated by us, that the methods could be generalized into this area. We do hope that this is so.

Cost-effectiveness of regulation

Regulation is supposed to contribute to better healthcare, and it is natural for Professor Bird to ask whether the process is cost effective compared with other potential uses of resources, and whether investment in better data would be worthwhile. This would be difficult to establish, as the consequences of different forms of regulation are hypothetical (although recent experience in the financial sector suggests that 'light touch' regulation can have serious unintended consequences!). She asks whether the higher qualification rate in risk-based inspections was just a consequence of the higher expected return when picking standards deliberately rather than at random. This may well be so but still shows that the identification of standards for inspection yields returns.

Professor Bird also wonders whether all this bureaucracy only identifies crises that were forewarned locally. We might expect this to be the case, since it is unlikely that any centralized monitoring mechanism would pick up issues of which there was no local knowledge. What is vital is to identify whether local proposals for improvement have been made and acted on.

Prospects

Many important questions await further research, such as Professor Bird's appeal to study the process of overdispersion—what is behind it and whether it is reducing. She also rightly emphasizes data quality, which Greg Phillpotts notes affects everything, not just outliers. Phillpotts also queries the extent to which our work takes the Royal Statistical Society's Working Party's recommendations forward (Bird *et al.*, 2005), whereas Tom King asks where are the indicators of 'success' rather than just 'lack of failure' and warns against constructing instruments used for political purposes and the need to improve transparency and accountability. We feel that we have gone some way in pushing forward the Working Party's agenda but admit that we have had limited influence in improving the quality of data for performance assessment, as the imperative not to add to data collection was strong.

We welcome the experience of Professor Mengersen and colleagues in Brisbane, who express scepticism about improvement arising from top-down quality initiatives. Traditional advice from the quality improvement 'industry' suggests that bottom up is better, with local ownership of data and responsibility for improvements. We strongly support their call for greater involvement in local projects—however, highly decentralized 'audit' has been rather discredited, and we see a role for a central resource for methods and central light touch scrutiny.

Finally, we are really grateful for Professor Woodall's comments and the connection to 'statistical engineering', which we admit was a new phrase for us. We would add to the list of *desiderata* that the statistical methods that are adopted should be as simple as possible, with robustness and transparency being more important than optimality under restricted assumptions. In this we follow Professor Militino and Professor Ugarte's recommendations that statistical procedures should be 'simple, clear and as intuitive as possible'. This means foregoing some sophistication in recognition of the low levels of numeracy that one is dealing with in many of these areas. We strongly agree that more of these 'statistical engineering' ideas should be promoted in education rather than just the simple single-data-set material that is taught in all statistics courses. But this will be difficult to teach and examine.

In conclusion, we are gratified by the response to the paper and hope that it stimulates further work in this area.

References in the discussion

Advisory Council on Mathematics Education (2011) *Mathematical Needs: Mathematics in the Workplace and in Higher Education*. London: Royal Society.

Assunção, R. and Correa, T. (2009) Surveillance to detect emerging space-time clusters. *Computnl Statist. Data Anal.*, **53**, 2817–2830.

Atkinson, A. C. and Riani, M. (2000) *Robust Diagnostic Regression Analysis*. New York: Springer.

Atkinson, A. C., Riani, M. and Cerioli, A. (2004) *Exploring Multivariate Data with the Forward Search*. New York: Springer.
Austin, P. C. (2009) Are (the log-odds of) hospital mortality rates normally distributed?: implications for studying variations in outcomes of medical care. *J. Evaln Clin. Pract.*, **15**, 514–523.
Baíllo, A. and Cuevas, A. (2006) Parametric versus nonparametric tolerance regions in detection problems. *Comput. Statist.*, **21**, 523–536.
Baíllo, A., Cuevas, A. and Justel, A. (2000) Set estimation and nonparametric detection. *Can. J. Statist.*, **28**, 765–782.
Bardsley, M., Sherlaw-Johnson, C., Blunt, I. and Spiegelhalter, D. J. (2009) Using routine intelligence to target inspection of healthcare providers in England. *Qual. Safty Hlth Care*, **18**, 189–194.
Benjamini, Y. and Kling, E. Y. (2007) The *p*-valued chart—a unified approach to statistical process control chart presentation. (Available from `businessken.co.uk/pvaluedSPC.aspx`.)
Benjamini, Y. and Kling, Y. (1999) A look at statistical process control through the *p*-values. *Technical Report RP-SOR-99-08*. Tel Aviv University, Tel Aviv.
Benning, A., Dixon-Woods, M., Nwulu, U., Ghaleb, M., Dawson, J., Barber, N., Franklin, B., Girling, A., Hemming, K., Carmalt, M., Rudge, G., Naicker, T., Kotecha, A., Derrington, C. and Lilford, R. (2011) Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase. *Br. Med. J.*, **342**, d199.
Benning, A., Ghaleb, M., Suokas, A., Dixon-Woods, M., Dawson, J., Barber, N., Franklin, B., Girling, A., Hemming, K., Carmalt, M., Rudge, G., Naicker, T., Nwulu, U., Choudhury, S. and Lilford, R. (2011) Large scale organizational intervention to improve patient safety in four UK hospitals: mixed method evaluation. *Br. Med. J.*, **342**, d195.
Berger, J. (2011) What are the open problems in Bayesian statistics? *Int. Soc. Baysn Anal. Bull.*, **18**, 1–4.
Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2011) Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *J. Am. Statist. Ass.*, to be published.
Bird, S. M. (2009) Buried data and the UK Healthcare Commission's legacy. *Lancet*, **373**, 1604–1605.
Bird, S. M., Cox, D., Farewell, V. T., Goldstein, H., Holt, T. and Smith, P. C. (2005) Performance indicators: good, bad, and ugly. *J. R. Statist. Soc.* A, **168**, 1–27.
Böhning, D., Malzahn, U., Dietz, E., Schlattmann, P., Viwatwongkasem, C. and Biggeri, A. (2002) Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics*, **3**, 445–457.
Böhning, D., Sarol, J., Rattanasiri, S., Viwatwongkasem, C. and Biggeri, A. (2004) A comparison of non-iterative and iterative estimators of heterogeneity variance for the standardized mortality ratio. *Biostatistics*, **5**, 61–74.
Boyne, G. A., James, O., John, P. and Petovsky, N. (2009) Democracy and Government Performance: holding incumbents to account in English local governments. *J. Polit.*, **71**, 1273–1284.
Campbell, M. J., Jacques, R. M., Fotheringham, J., Pearson, T., Maheswaran, R. and Nicholl, J. (2011) An evaluation of the Summary Hospital Mortality Index. *Final Report*. School of Health and Related Research, University of Sheffield, Sheffield. (Available from `http://www.sheffield.ac.uk/scharr/sections/hsr/statistics`.)
Carthey, J., de Leval, M. R. and Reason, J. T. (2001) The human factor in cardiac surgery: errors and near misses in a high technology medical domain. *Ann. Thorac. Surg.*, **72**, 300–305.
DeGroot, M. H. (1970) *Optimal Statistical Decisions*. New York: McGraw-Hill.
De Martini, D. (2008) Reproducibility probability estimation for testing statistical hypotheses. *Statist. Probab. Lett.*, **78**, 1056–1061.
Deming, W. E. (2000) *Out of the Crisis*. Cambridge: MIT Press.
Devroye, L. and Wise, G. (1980) Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, **38**, 480–488.
Diggle, P. J., Rowlingson, B. and Su, T. L. (2005) Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, **16**, 423–434.
Duval, S. and Tweedie, R. (2000a) Trim and fill: a simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**, 276–284.
Duval, S. and Tweedie, R. (2000b) A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *J. Am. Statist. Ass.*, **95**, 89–98.
Efron, B. (2008) Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, **23**, 1–22.
Frisén, M. (2003) Statistical surveillance: optimality and methods. *Int. Statist. Rev.*, **71**, 403–434.
Gandy, A. and Kvaløy, J. (2011) Guaranteed conditional performance of control charts via bootstrap methods. *Preprint*.
Gelman, A. (2009) Columbia University workshop on Gelman, A., Hill, J. and Yajima, M. (2009): Why we (usually) don't have to worry about multiple comparisons. *Minutes 12–13*. (Available from `http://www.stat.columbia.edu/~martin/Workshop/statistics_neuro_data_931_speaker_04.mov`.)
Gelman, A. and Price, P. (1999) All maps of parameter estimates are misleading. *Statist. Med.*, **18**, 3221–3234.
Goldstein, H. and Speigelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of performance (with discussion). *J. R. Statist. Soc.* A, **159**, 385–443.

Goodman, S. N. (1992) A comment on replication, p-values and evidence. *Statist. Med.*, **11**, 875–879.
Hampton, P. (2005) Reducing administrative burdens: effective inspection and enforcement. (Available from `http://www.berr.gov.uk/whatwedo/bre/inspection-enforcement/assessing-regulatory-system/page44042.html#`.)
Hawkins, D. M. (1993) Robustification of cumulative sum charts by Winsorization. *J. Qual. Technol.*, **25**, 248–261.
Hoerl, R. W. and Snee, R. D. (2010) Moving the statistics profession forward to the next level. *Am. Statistn*, **64**, 10–14.
Houstein, T., Gastmeier, P., Holmes, A., Lucet, J., Shannon, R., Pittet, D. and Harbath, S. (2011) Use of benchmarking and public reporting for infection control in four high-income countries. *Lancet Infect. Dis.*, **11**, 471–481.
Huskins, C., Huckabee, C., O'Grady, N., Murray, P., Kopetskie, H., Zimmer, L., Walker, M., Sinkowitz-Cochran, R., Jernigan, J., Samore, M., Wallace, D. and Goldmann, D. for the STAR*ICU Trial Investigators (2011) Intervention to reduce transmission of resistant bacteria in intensive care. *New Engl. J. Med.*, **364**, 1407–1418.
Jani, R., Kralovic, S., Evans, M., Ambrose, M., Simbartl, L., Obrovsky, D., Render, M., Freyberg, R., Jernigan, J., Muder, R., Miller, L. and Roselle, G. (2011) Methicillin-resistant Staphylococcus aureus infections. *New Engl. J. Med.*, **364**, 1419–1430.
Johnson, N. (2007) *Simply Complexity*. Oxford: Oxford Oneworld.
Jones, H. E. and Spiegelhalter, D. J. (2011) The identification of 'unusual' health-care providers from a hierarchical model. *Am. Statistn*, to be published.
King, T. (2011) Statistics in Society: three case studies in the UK. *Applications and Policy Working Paper A11/01*. Southampton Statistical Sciences Research Institute, University of Southampton, Southampton. (Available from `http://www.soton.ac.uk/s3ri/publications/details.php?id=170`.)
Kipnis, P., Escobar, G. J. and Draper, D. (2010) Effect of choice of estimation method on inter-hospital mortality rate comparisons. *Med. Care*, **48**, 458–465.
Lambert, D. and Liu, C. (2006) Adaptive thresholds: monitoring streams of network counts. *J. Am. Statist. Ass.*, **101**, 78–88.
Landrigan, C., Parry, G., Bones, C., Hackbarth, A., Goldmann, D. and Sharek, P. (2010) Temporal trends in rates of patient harm resulting from medical care. *New Engl. J. Med.*, **363**, 2124–2134.
Lawson, A. B. and Kleinman, K. (2005) *Spatial and Syndromic Surveillance for Public Health*. New York: Wiley.
Lin, R., Louis, T. A., Paddock, S. and Ridgeway, G. (2006) Loss function based ranking in two-stage, hierarchical models. *Baysn Anal.*, **1**, 915–946.
Lin, R., Louis, T. A., Paddock, S. and Ridgeway, G. (2009) Ranking USRDS, provider-specific SMRs from 1998–2001. *Hlth Serv. Outcms Res. Method.*, **9**, 22–38.
Lindley, D. (1997) Review of *Multiple Comparisons: Theory and Methods*, by J. C. Hsu. *Statistician*, **46**, 572–573.
Lindley, D. V. (1985) *Making Decisions*, 2nd edn. New York: Wiley.
Lindley, D. V. (1998) Decision analysis and bioequivalence trials. *Statist. Sci.*, **13**, 136–141.
Liu, R. Y. (1995) Control charts for multivariate processes. *J. Am. Statist. Ass.*, **90**, 1380–1387.
Liu, R. Y. and Singh, K. (1993) A quality index based on data depth and multivariate rank tests. *J. Am. Statist. Ass.*, **88**, 252–260.
Lockwood, J. R., Louis, T. A. and McCaffrey, D. (2002) Uncertainty in rank estimation: implications for value added modeling accountability systems. *J. Educ. Behav. Statist.*, **27**, 255–270.
Longford, N. T. (2005) Model selection and efficiency—is 'Which model...?' the right question? *J. R. Statist. Soc.* A, **168**, 469–472.
Longford, N. T. (2010) Bayesian decision making about small binomial rates with uncertainty about the prior. *Am. Statistn*, **64**, 164–169.
Longford, N. T. (2011) Comparing normal random samples, with uncertainty about the priors and utilities. *Scand. J. Statist.*, **38**, to be published.
Louis, T. A. (1984) Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Am. Statist. Ass.*, **78**, 393–398.
Millar, A. (2011) Is money spent on quality improvement better spent on clinical care? *Med. J. Aust.*, **194**, 640.
Morton, A. (2011) Hospital safety and complexity. *Br. Med. J.*, **342**, d1320.
National Audit Office (2001) Inappropriate adjustments to NHS waiting lists. *Press Release*, Dec. 19th.
Normand, S.-L. T. and Shahian, D. M. (2007) Statistical and clinical aspects of hospital outcomes profiling. *Statist. Sci.*, **22**, 206–226.
Paddock, S. M. and Louis, T. A. (2011) Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *Appl. Statist.*, **60**, 575–589.
Powell, G., Caan, W. and Crowe, M. (1994) What events precede violent incidents in psychiatric hospitals? *Br. J. Psychiatr.*, **165**, 107–112.
Pratt, R. (2011) Time for a culture change. *New Engl. J. Med.*, **364**, 1464–1465.
Pronovost, P., Berenholtz, S. and Morlock, L. (2011) Is quality of care improving in the UK? *Br. Med. J.*, **342**, c6646.

Rascz, M. J. and Sedransk, J. (2010) Bayesian and frequentist methods for provider profiling using risk-adjusted assessments of medical outcomes. *J. Am. Statist. Ass.*, **105**, 48–58.
Riani, M., Atkinson, A. C. and Cerioli, A. (2009) Finding an unknown number of multivariate outliers. *J. R. Statist. Soc.* B, **71**, 447–466.
Riani, M., Cerioli, A., Atkinson, A., Perrotta, D. and Torti, F. (2008) Fitting mixtures of regression lines with the forward search. In *Mining Massive Data Sets for Security* (eds F. Fogelman-Soulié, D. Perrotta, J. Piskorski, and R. Steinberger), pp. 271–286. Amsterdam: IOS.
Runciman, W. (2011) Is money spent on quality improvement better spent on clinical care? *Med. J. Aust.*, **194**, 641.
Scottish Environment Protection Agency (2011) Water classification reports for 2007, 2008 and 2009. Scottish Environment Protection Agency, Stirling. (Available from `http://www.sepa.org.uk/water/monitoring-and-classification.aspx`.)
Shao, J. and Chow, S. C. (2002) Reproducibility probability in clinical trials. *Statist. Med.*, **21**, 1727–1742.
Simpson, S. and Dorling, D. (1999) Conclusion: Statistics and 'the truth'. In *Statistics in Society* (eds D. Dorling and S. Simpson), pp. 414–420. London: Arnold.
Smith, I., Rivers, J., Mengersen, K. and Cameron, J. (2010) Performance monitoring in interventional cardiology: application of statistical process control to a single-site database. *EuroIntervention*, **6**, 955–962.
Snee, R. D. and Hoerl, R. W. (2011) Proper blending—the right mix between statistical engineering, applied statistics. *Qual. Prog.*, **44**, 46–49.
Sparrow, M. K. (2000) *The Regulatory Craft: Controlling Risks, Solving Problems, and Managing Compliance*. Harrisonburg: Donnelley.
Spiegelhalter, D. (2005) Problems in assessing rates of infection for methicillin resistant *Staphylococcus Aureus*. *Br. Med. J.*, **331**, 1013–1015.
Spiegelhalter, D. J., Aylin, P., Best, N. G., Evans, S. J. W. and Murray, G. D. (2002) Commissioned analysis of surgical performance using routine data: lessons from the Bristol inquiry (with discussion). *J. R. Statist. Soc.* A, **165**, 191–231.
Vincent, C. (2010) *Patient Safety*, 2nd edn. London: Wiley–Blackwell–BMJ Books.
Waterhouse, M., Morton, A., Mengersen, K., Cook, D. and Playford, G. (2011) Role of overcrowding in methicillin-resistant *Staphylococcus Aureus* transmission: Bayesian network analysis for a single public hospital. *J. Hosp. Infectn*, 92–96.
Williams, D. (2011) More performance data to be revealed to public. *Hlth Serv. J.*, June 2nd, 7.
Young, I. M. (2000) Inclusive political communication. In *Inclusion and Democracy* (ed. I. M. Young), pp. 52–80. Oxford: Oxford University Press.