

Funnel plots for comparing institutional performance

David J. Spiegelhalter^{*,†}

MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, U.K.

SUMMARY

‘Funnel plots’ are recommended as a graphical aid for institutional comparisons, in which an estimate of an underlying quantity is plotted against an interpretable measure of its precision. ‘Control limits’ form a funnel around the target outcome, in a close analogy to standard Shewhart control charts. Examples are given for comparing proportions and changes in rates, assessing association between outcome and volume of cases, and dealing with over-dispersion due to unmeasured risk factors. We conclude that funnel plots are flexible, attractively simple, and avoid spurious ranking of institutions into ‘league tables’. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: control charts; outliers; over-dispersion; institutional profiling; ranking

1. INTRODUCTION

Demands for increased accountability of public services have led to increased attention to institutional comparisons on the basis of quantitative outcome measures, whether school examination results, surgical mortality rates, or research output from universities. Here ‘institution’ refers to any unit of analysis, which in the health context could be a health authority, hospital, surgical team or even an individual named surgeon. Such comparisons commonly lead to the production of ‘league tables’, in which institutions are ranked according to a performance indicator and, possibly with the aid of confidence intervals, ‘outlying’ institutions identified. For example, Figure 1 shows a league table of hospitals based on mortality following a fractured hip—this display is similar to that of the original publication [1].

Such presentations have been criticized as leading to a spurious focus on rank ordering, when it is known that the rank of an institution is one of the most difficult quantities to estimate [2, 3]. Mohammed *et al.* [4] argued strongly that a more appropriate presentation would be based on Shewhart’s control charts [5], in which ‘in-control’ institutions are assumed to be subject to ‘common-cause’ variability, whereas those that are ‘out-of-control’ will exhibit

*Correspondence to: David J. Spiegelhalter, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.

†E-mail: david.spiegelhalter@mrc-bsu.cam.ac.uk

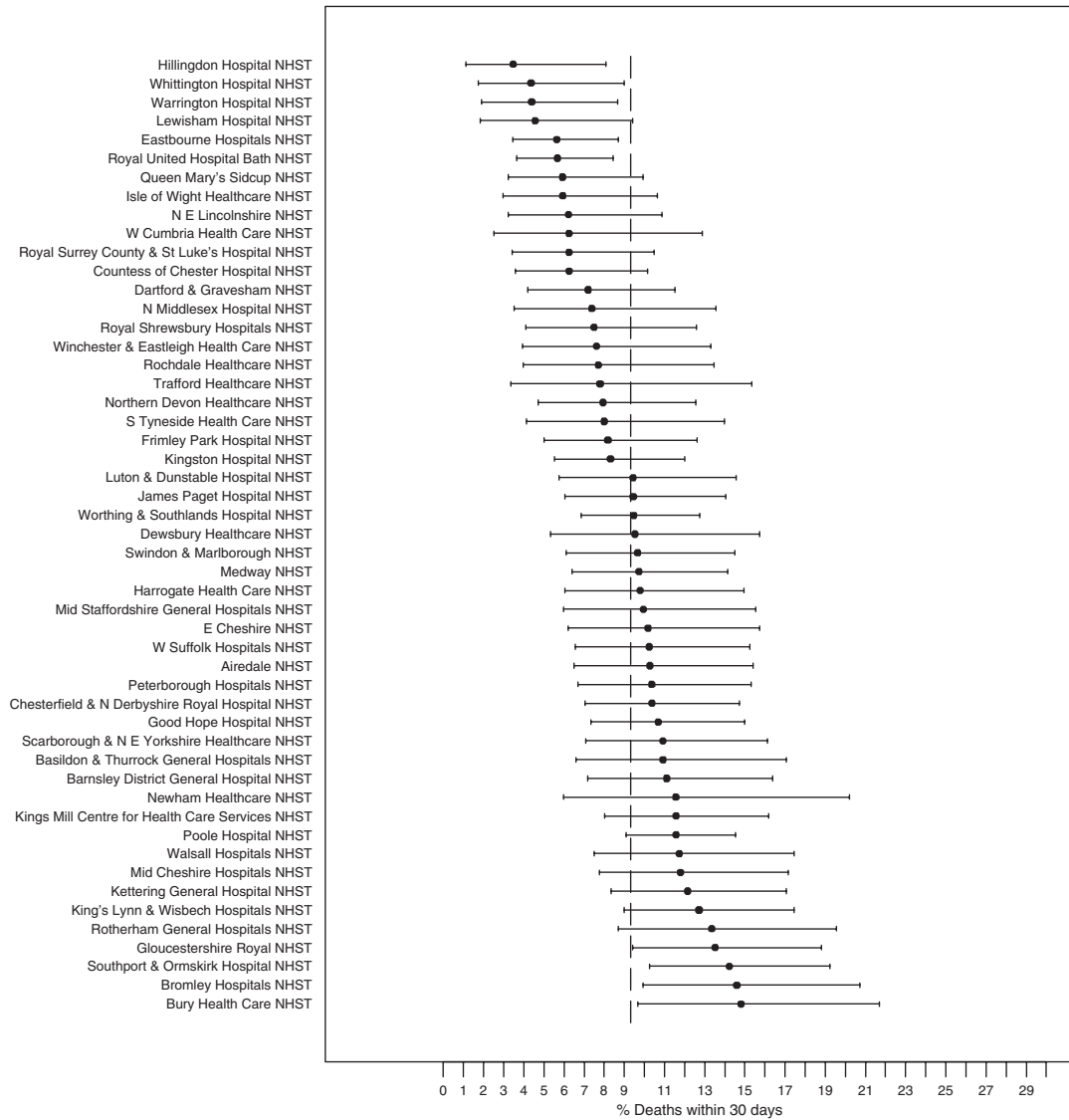


Figure 1. ‘Caterpillar’ plot of 30-day mortality rates, age and sex standardized, following treatment for fractured hip for over-65’s in 51 medium acute and multi-service hospitals in England, 2000–2001. Ninety-five per cent confidence intervals are plotted and compared to the overall proportion of 9.3 per cent.

‘special cause’ variability: a threshold of 3 standard deviations being commonly used as a demarcation between these two categories.

In this paper we argue that a suitable form of such a control chart is the ‘funnel plot’ in which the observed indicator is plotted against a measure of its precision, so that the control limits form a ‘funnel’ around the target outcome. For example, Figure 2 shows the data pre-

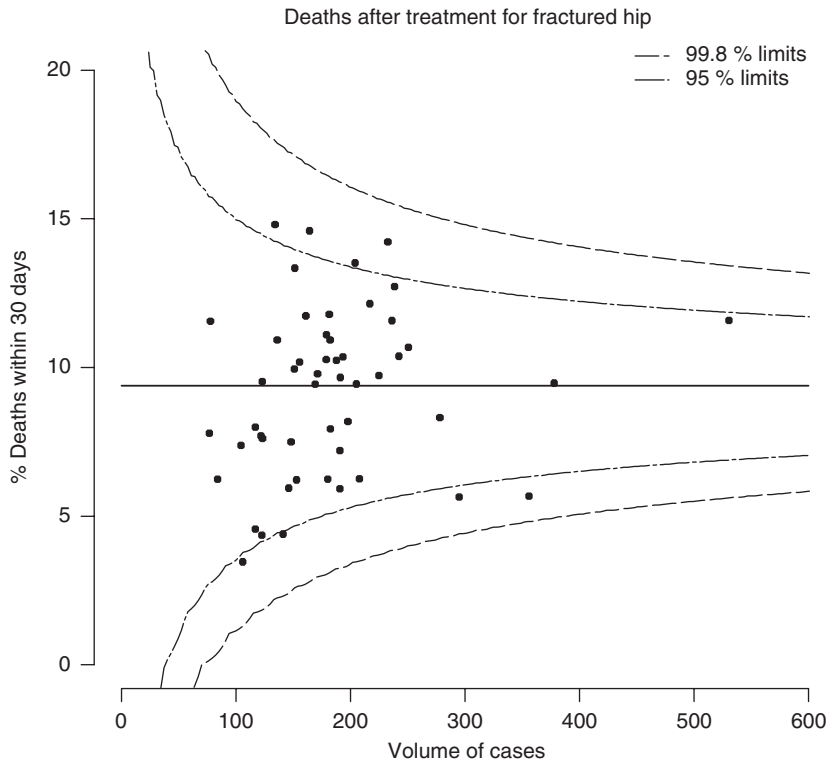


Figure 2. Funnel plot of 30-day age–sex standardized mortality rates following treatment for fractured hip of over-65's in 51 medium acute and multi-service hospitals in England, 2000–2001. The target is the overall proportion of 9.3 per cent. (Any roughness of the plotted control limits is due to interpolation between exact Binomial limits.)

sented in Figure 1, but now plotting the observed rate against the volume and superimposing 95 per cent (≈ 2 standard deviation) and 99.8 per cent (≈ 3 standard deviation) prediction limits around the overall mortality rate. This funnel plot clearly reveals the bulk of the institutions as lying within the 95 per cent limits, and in particular emphasises that there is no basis for ranking the hospitals. Four hospitals might perhaps be further scrutinized as having high rates, but no hospital lies outside the 99.8 per cent limits. Such a plot suggests there may be little to be learnt by seeking to understand reasons for the variability, although it may be worth pooling data over a longer period to see if there was evidence of consistent differences in outcome.

Such plots are not novel: they are a standard tool within meta-analysis as a graphical check of any relationship between effect estimates and their precision that might suggest publication bias [6], and they have also been previously been used for comparing clinical outcomes [7–11]. Here we add to this development by providing full details for their construction, assessing volume/outcome association and dealing with 'over-dispersion'.

Section 2 provides the formal definition for a funnel plot, and examples for cross-sectional proportions, risk-adjusted rates, and changes in rates are given in Section 3. Section 4 illustrates how the plots can be used to assess the relationship between outcomes and volume of

cases, and Section 5 illustrates how one might extend the method to deal with over-dispersion due to unmeasured risk factors leading to unexplained but acceptable variability between in-control units. Some brief conclusions are drawn in Section 6. Appendix A provides exact and approximate formulae for proportions, rates (whether indirectly or directly standardized), and continuous responses, both for cross-sectional data and changes.

2. FUNNEL PLOTS

A funnel plot has four components:

1. An *indicator* Y .
2. A *target* θ_0 for Y which specifies the desired expectation, so that $\mathbb{E}(Y|\theta_0) = \theta_0$ for those institutions considered 'in-control'.
3. A *precision* parameter ρ that determines the accuracy with which the indicator is being measured, so that the 'null distribution' of Y , given an in-control institution in which the target is being achieved, is taken as $p(y|\theta_0, \rho)$. In the contexts we consider ρ is taken as (possibly approximately) proportional to the inverse variance $1/\mathbb{V}(Y|\theta_0)$, i.e.

$$\rho = g(\theta_0)/\mathbb{V}(Y|\theta_0) \quad (1)$$

for some function g . There is some degree of arbitrariness about the choice of ρ , although it is best to select a measure which is directly interpretable. For example, when Y is a proportion we have $\mathbb{V}(Y|\theta_0) = \theta_0(1 - \theta_0)/n$, and since an interpretable measure of precision is the sample size it is natural to set so $\rho = n$ and hence take $g(\theta_0) = \theta_0(1 - \theta_0)$. Funnel plots for changes in performance may be more complex, with ρ depending on a nuisance parameter that needs to be estimated.

4. *Control limits* $y_p(\theta_0, \rho)$ for a P -value p , where the chance of exceeding these limits for an in-control unit is p . Hence $y_p = F^{-1}(p)$, where F^{-1} is the inverse cumulative distribution function of $p(y|\theta_0, \rho)$, so that $F(y_p|\theta_0, \rho) = p$. $p = 0.001, 0.999$ and $p = 0.025, 0.975$ may be reasonable standards corresponding approximately to 2 and 3 standard deviation intervals, with the latter corresponding to the classic Shewhart limits [5].

Given a series of I observations y_i with associated precisions ρ_i , a funnel plot then consists of a plot of y_i against ρ_i , with target θ_0 shown by a horizontal line, superimposed on the control limits plotted as a function of ρ . Note that the control limits can be 'pre-drawn' as they do not depend on the data being plotted. Also the axes may need to be transformed or relabelled to improve interpretability: for example it may be convenient to create control limits on a logarithmic scale, but axes should be labelled on the natural scale.

In many circumstances we can assume an exact or approximate normal distribution

$$Y|\theta_0, \rho \sim N[\theta_0, g(\theta_0)/\rho] \quad (2)$$

Control limits are then plotted at

$$y_p(\theta_0, \rho) = \theta_0 + z_p \sqrt{g(\theta_0)/\rho} \quad (3)$$

where z_p is such that $\Phi(z_p) = P(Z \leq z_p) = p$ for a standard normal variate Z , so that, for example, $z_{0.025} = -1.96$. Plotted points will, for example, lie above the p control limits if

and only if $y_i \geq y_p$, equivalent to $(y_i - \theta_0)\sqrt{\rho_i/g(\theta_0)} \geq z_p$. Here $(y_i - \theta_0)\sqrt{\rho_i/g(\theta_0)}$ is the standardized Pearson residual

$$z_i = \frac{y_i - \theta_0}{\sqrt{\mathbb{V}(Y|\theta_0)}} \quad (4)$$

we shall also term this the ‘naive’ Z -score for the i th unit.

For discrete data it is not possible to invert the cumulative distribution function F exactly, and hence some interpolation will be necessary to establish the control limits: see Sections A.1.1 and A.2.1 for details.

We note that our choice of ρ as proportional to the inverse variance does not match the recommendations made by Sterne and Egger [12] in the context of detecting publication bias in meta-analysis, who suggest plotting outcome against standard error so that the control limits are straight lines. As disadvantages of the inverse variance, they cite the curved control limits and compression of smaller studies at the bottom of the funnel when there are some extremely large studies. We feel that these are not vital issues in the context of institutional comparisons, and are superseded by ready interpretability of the precision measure as being, for example, the sample size.

3. EXAMPLES

3.1. Risk-adjusted data

Risk-adjustment procedures for clinical outcomes generally take the form of indirect standardization: a logistic regression model is used to derive a probability of an event (say death) for individuals with a particular set of covariates, and these can be summed over the set of patients being treated by an institution or an individual to give an expected number of events E . This can be contrasted with the observed number of events O to give a standardized event ratio O/E or difference $O - E$.

As an example of this procedure, we consider data from the New York State Coronary Artery Bypass Graft (CABG) programme, in which all such operations in New York State are recorded and the outcomes published by named hospital and surgeon using 3-year moving totals. By multiplying the standardized mortality ratio $SMR = O/E$ by the overall state-wide mortality rate (2.2 per cent in 1997–1999) a risk-adjusted mortality rate is obtained. Figure 3 shows funnel plots by hospital and surgeon using publicly available data [13].

The funnel plot by hospital clearly identifies an interesting high-mortality low-volume hospital, as well as a borderline low-mortality high-volume hospital. There is a suggestion of an outcome-volume association which is further explored in Section 4. The great majority of the individual surgeons follow the pattern delineated by the funnel, but there are nine who are outside the upper 99.8 per cent limits. Current policy is that those individuals whose 95 per cent intervals exclude the state-wide average are ‘starred’ in the publication: there were 19 occurrences of this (12 with high rates, 7 with low) compared to 8.3 expected by chance in 175 instances, illustrating the problems of multiple testing that arise in this context (Section 6).

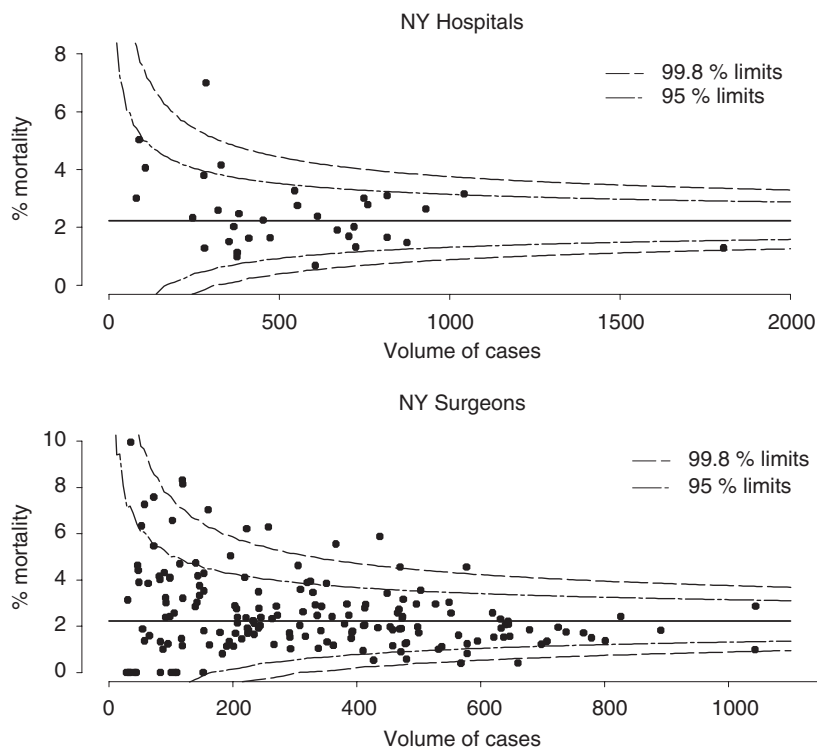


Figure 3. Risk-adjusted 30-day mortality rates following coronary artery bypass grafts in New York state, 1997–1999, for 33 hospitals and 175 surgeons who conducted at least 25 operations in a hospital (separate results are given for the same surgeon operating in different hospitals, and some comprise an ‘all other’ category). The target is the overall rate of 2.2 per cent.

3.2. Change data

A variety of measures of change in performance are possible. For example, changes in proportions may be summarized by absolute differences, risk ratios or odds ratios, and in Appendix A we suggest an appropriate precision measure ρ in each instance. Here we only consider change between performance indicators observed at two periods: the additional problems of multiplicity that occur if sequential monitoring is being carried out are discussed briefly in Section 6.

The NHS Plan for the U.K. [14] set a target that by 2004 the rate of pregnancies in under 18-year-olds would be reduced by 15 per cent from its 1998 level. The number and population rates of teenage pregnancy, defined as the rate of live births, still births and abortions per 1000 women aged 15–17, are routinely collected by the Teenage Pregnancy Unit, and the ratio of 2001 to 1998 rates are plotted in Figure 4 by top-tier health authority in England. These could be treated as ratios of directly standardized rates and plotted against the underlying population denominator, but here we have considered the rates as indirectly standardized assuming a constant population and used the exact limits based on the conditional distribution

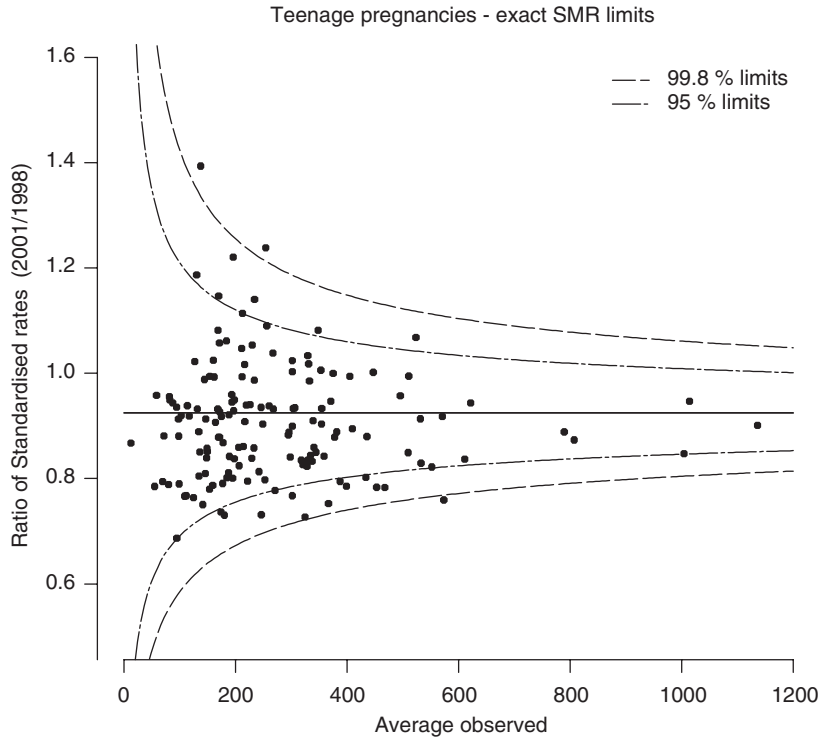


Figure 4. Ratios of teenage-pregnancy rates (2001/1998) for 148 top-tier health authorities in England. A government target of a 7.5 per cent reduction (ratio = 0.925) had been set. Control limits are based on exact Binomial limits using a conditional argument.

(see Section A.2.2 for details). The 1998–2001 target of 0.925 (7.5 per cent reduction) is used: the overall observed ratio is 0.90 (10 per cent reduction).

It is remarkable how well the authorities have met the target set. Only a few health authorities exhibit variability of particular interest, and attention should perhaps focus on those three lying outside the 99.8 per cent limits. The plot emphasises that there is no point in further investigation of the many authorities whose rates have increased but who still lie within the funnel—cases of this are only to be expected.

4. RELATIONSHIP WITH VOLUME

It is natural to use funnel plots as an initial assessment of the relationship between outcome and precision, and indeed this is their main use in the context of meta-analysis. For example, Figure 5 shows mortality rates following paediatric cardiac surgery in 12 English hospitals between 1991 and 1995 [9, 15]. The most obvious feature is the strongly divergent performance of one hospital (Bristol Royal Infirmary), but even disregarding this institution there is a suggestion of an association between outcome and volume.

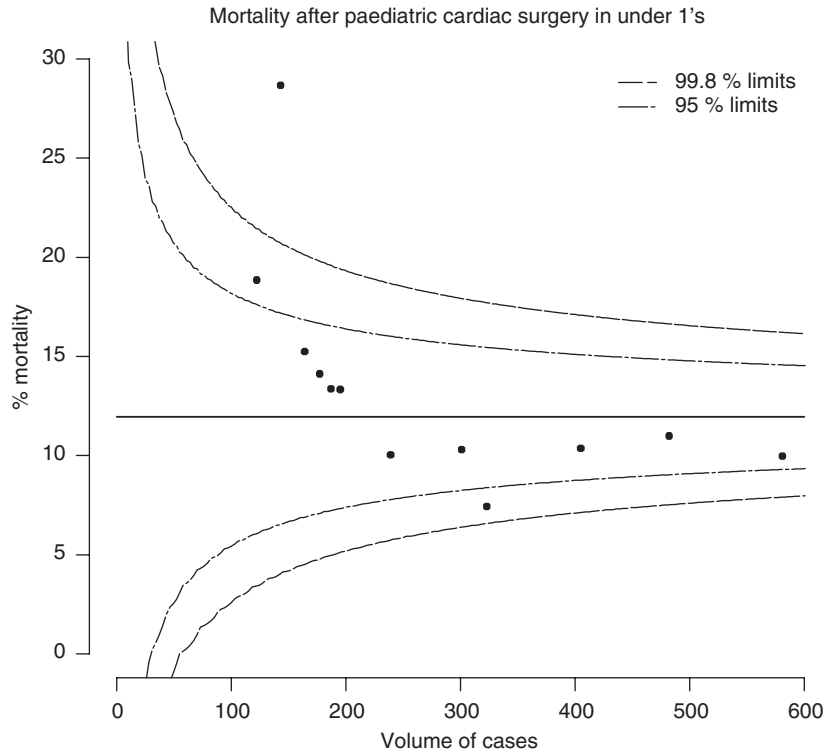


Figure 5. Mortality rates following paediatric cardiac surgery in under-1's in 12 English specialist centres, using Hospital Episode Statistics 1991–1995. The target is the overall average rate of 12.0 per cent.

A formal test of this association can be obtained by a regression of y_i on ρ_i (or a function of ρ_i) using a model with an appropriate error structure. For example, for the paediatric cardiac surgery data we assume in hospital i that there are r_i deaths out of n_i patients. A suitable model might be

$$r_i \sim \text{Binomial}(p_i, n_i)$$

$$\log \frac{p_i}{1 - p_i} = \alpha + \beta \log n_i \quad (5)$$

The coefficient β in (5) then has the following attractive interpretation: a small percentage rise in sample size, say 10 per cent, leads approximately to a $\beta \times 10$ per cent change in the odds of death. Excluding Bristol from the analysis we estimate $\hat{\beta} = -0.37$, 95 per cent interval -0.60 to -0.13 , 2-sided $P = 0.002$, so that a 10 per cent increase in volume is estimated to be associated with a 4 per cent relative decrease in mortality. An association with volume is also found for the New York data in Section 3.1, with a 10 per cent increase in volume being associated with a 2.7 per cent ($2P = 0.003$) decrease in hospital mortality, and a 2.7 per cent ($2P < 0.001$) decrease in surgeon mortality. The single low-mortality high-volume hospital and surgeon are clearly influential in this assessment.

Problems arise if ρ_i depends on a nuisance parameter and hence is not known exactly, since then we have the well-known problem of carrying out a regression when both dependent and independent variables are measured with (possibly correlated) error. Irwig *et al.* [16] identified the potential bias arising from this in meta-analytic applications where the effect estimate and its standard error can be highly correlated even when there is no true association between effect and precision, and it has been suggested [17] that asymmetry in many funnel plots could be due to this bias. Appropriate analysis methods may involve more complex likelihood or Bayesian methods that are beyond the scope of this paper.

5. OVER-DISPERSION

So far the funnel plots considered have been based on the assumption that the null (target) distribution fully expresses the variability of the in-control units, but in many situations this assumption will not hold and even units that are judged not to be special cases will be ‘over-dispersed’ around the target. For example, Figure 6(a) shows rates of emergency readmissions within 28-days of discharge for 67 hospitals in England in 2000–2001 [1]. The large majority of institutions lie outside the funnel, casting doubt on the appropriateness of the limits. The original presentation was as a ranked league table highlighting all hospitals which were ‘significantly different’ from the national average, leading to the anomalous situation of the majority of hospitals being labelled as ‘divergent’.

In general this behaviour will be due to the impact of unmeasured covariates that are not taken into account in any risk-adjustment method: although each may have a small impact on the outcome and be insufficient to merit a claim of ‘out-of-control’, when taken together they can lead to an excess variability among in-control units [18]. This is likely to become particularly apparent in high-volume outcome measures such as that of Figure 6. One can think of this as ‘acceptable’ excess variation, and if it is not taken into account there will be an inappropriate number of units identified as special cases by the funnel plot. It may best be a matter of judgement where the boundary between ‘acceptable’ and ‘out-of-control’ lies, but an automated default procedure is clearly useful.

We assumed in Section 2 that $\rho = g(\theta_0)/\mathbb{V}_0(Y|\theta_0, \rho)$, where the subscript on \mathbb{V}_0 has now been introduced to indicate this is the situation with no allowance for over-dispersion. We shall explore two basic statistical models for over-dispersion: first the standard approach used in generalized linear modelling [19], and then a ‘random-effects’ formulation which adds a constant term to the sampling variance of each unit.

The former is a multiplicative approach that introduces an over-dispersion factor ϕ that will inflate the null variance, so that

$$\mathbb{V}(Y|\theta_0, \rho, \phi) = \phi \mathbb{V}_0(Y|\theta_0, \rho) = \frac{\phi g(\theta_0)}{\rho}$$

Suppose we have a sample of I units that we shall assume (for the present) all to be in-control. ϕ may be estimated as follows [19]:

$$\hat{\phi} = \frac{1}{I} \sum_i \frac{(y_i - \theta_0)^2 \rho_i}{g(\theta_0)} = \frac{1}{I} \sum_i z_i^2 \quad (6)$$

where z_i is the standardized Pearson residual defined in (4).

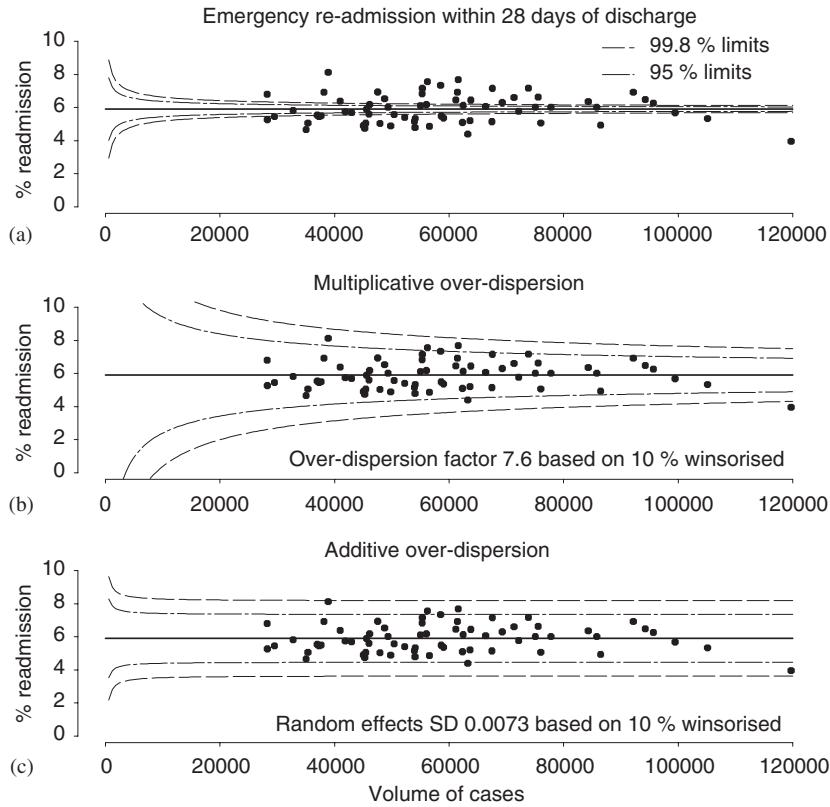


Figure 6. Proportions of emergency re-admission within 28 days of discharge from 67 large acute or multi-service hospitals in England, 2000–2001. The target is the overall average rate of 5.9 per cent: (a) shows the over-dispersion around the unadjusted funnel plot; (b) is a multiplicative model in which the limits are expanded by an estimated over-dispersion parameter ϕ , while (c) is an additive random-effects model in which all the sample variances have an additional τ^2 term added.

The current control limits can then be inflated by a factor $\sqrt{\hat{\phi}}$ around θ_0 . For example, based on the approximate normal control limits, over-dispersed control limits can then be plotted as

$$y_p(\theta_0, \rho) = \theta_0 + z_p \sqrt{\hat{\phi} g(\theta_0) / \rho} \tag{7}$$

equivalent to creating a ‘modified’ Z-score $z_i / \sqrt{\hat{\phi}}$ and comparing to standard normal deviates.

As emphasized by Marshall *et al.* [18], the problem is identifying the in-control units on which to base the estimate of ϕ . There is an element of circularity, in that if out-of-control units are included in this estimation process, they will tend to increase the estimate of ϕ , widen the funnel limits and hence make it more difficult to detect the very cases in which we are interested. Therefore when estimating ϕ , we may want to ‘robustify’ the analysis by

minimizing the influence of outlying cases that the system is designed to detect. Many options exist: here we propose a simple ‘Winsorised’ estimate in which the most extreme cases are shrunk to pre-specified percentiles of the distribution.

1. Rank cases according to their naive Z -scores.
2. Identify Z_q and Z_{1-q} , the $100q$ per cent most extreme top and bottom naive Z -scores, where q might, for example, be 0.1.
3. Set the lowest $100q$ per cent of Z -scores to Z_q , and the highest $100q$ per cent of Z -scores to Z_{1-q} . Denote the resulting set of Z -scores, both those left unchanged and those that have been ‘pulled-in’, by Z^W .
4. Calculate the estimate $\hat{\phi}$ in (6) using Z^W , so that

$$\hat{\phi}^W = \frac{1}{I} \sum_i Z_i^W(q)^2 \tag{8}$$

The estimate $\hat{\phi}^W$ might additionally be multiplied by a debiasing factor $w(q)$ based on the fact that, if all the institutions are in-control and the only variability is due to over-dispersion, the variance of the Winsorized Z -statistics will tend to be less than ϕ . Specifically, if $Z^W(q)$ arises from a q -Winsorized standard normal distribution, then its variance can be shown to be

$$\mathbb{V}[Z^W(q)] = 1 + 2q(z_q^2 - 1) - 2z_q e^{-z_q^2/2} / \sqrt{2\pi} = 1/w(q) \tag{9}$$

where $z_q = \Phi^{-1}(1 - q)$. Specific values include $w(0.05) = 1.20$, $w(0.10) = 1.47$, *e.g.* 10 per cent Winsorization will lead to the variance being $1/1.47 = 0.68$ of its true value, and hence will underestimate the variance by around 32 per cent. We have found that using this factor leads to somewhat wide limits and hence have not adopted it in the example below.

If there is no true over-dispersion, then $I\hat{\phi}$ has approximately a χ^2_I distribution, which means that $\mathbb{E}(\hat{\phi}) = 1$, $\mathbb{V}(\hat{\phi}) = 2/I$. Rather than applying the over-dispersion adjustment to all data by default, it may therefore be better to:

1. not assume *under*-dispersion: i.e. if $\hat{\phi} < 1$, assume $\phi = 1$;
2. demand a ‘statistically significant’ $\hat{\phi}$ before including an adjustment for over-dispersion: i.e. assume $\phi = 1$ unless the estimated $\hat{\phi} > 1 + 2\sqrt{2/I}$.

Returning to the example, we see in Figure 6(b) that allowance for multiplicative over-dispersion leads to the width of the control limits being multiplied by an over-dispersion factor $\sqrt{\hat{\phi}} = 7.6$ based on 10 per cent Winsorization. The observations follow the funnel reasonably well and only one (very large) hospital stands out as having a particularly low rate.

The second, random-effects approach assumes that Y_i has expectation $\mathbb{E}(Y_i) = \theta_i$ and variance $\mathbb{V}(Y_i) = s_i^2$, and that for ‘on-target’ trusts θ_i is distributed with mean θ_0 and standard deviation τ . Hence the null hypothesis is a distribution rather than a point. τ can be estimated using a standard ‘method of moments’ [20] estimator

$$\hat{\tau}^2 = \frac{I\hat{\phi} - (I - 1)}{\sum_i w_i - \sum_k w_k^2 / \sum_i w_i} \tag{10}$$

where $w_i = 1/s_i^2$, and $\hat{\phi}$ is the test for heterogeneity: if $\hat{\phi} < (I - 1)/I$, then $\hat{\tau}^2$ is set to 0 and complete homogeneity is assumed. The funnel plot boundaries are then given by

$$\theta_0 \pm z_p \sqrt{\mathbb{V}(Y|\theta_0, \rho) + \tau^2} \quad (11)$$

Figure 6(c) shows that for the re-admission data τ is estimated to be 0.73 per cent, and the funnel is essentially independent of volume and appears to fit the data well. This additive random-effects formulation is to be used by the Healthcare Commission for England and Wales when dealing with possible over-dispersion in the performance indicators contributing to the current 'star ratings' procedure [21].

Adjustment for over-dispersion should, however, be seen as purely a temporary measure to get reasonable control limits given the current state of risk-adjustment, and attempts should clearly be made to explore reasons for the excess variability and either redefine the indicator or adjust for additional risk factors. Indicators such as that displayed in Figure 6 are clearly based on a very heterogeneous set of outcomes and it would be unfortunate if a statistical 'fix' were seen as an alternative to more careful exploration of an appropriate performance measure.

6. CONCLUSIONS

In our personal experience, funnel plots are very attractive to consumers of data on institutional comparisons. Advantages include:

- The axes are readily interpretable, so that additional observations can be added by hand if desired.
- The eye is naturally drawn to important points that lie outside the funnels.
- There is no spurious ranking of institutions.
- There is clear allowance for additional variability in institutions with small volume.
- The relationship of outcome with volume can be both informally and formally assessed.
- Over-dispersion can be taken into account.
- If repeated observations are made over time then it would be possible to plot repeated points and join them up to show progress (although see the comment below).
- They are easy to produce within popular spreadsheet programmes.

However, some limitations should be acknowledged:

- The only allowance for multiple comparisons is the use of small P -values for the control limits. These could be chosen based on some formal criterion such as Bonferroni or fixing the False Discovery Rate [22], but we suggest this is best carried out separately.
- For change measures there is no formal allowance for 'regression to the mean', in which extreme outcomes are expected to tend towards the population mean simply because a contributing factor to their 'extremeness' is likely to be a run of good or bad luck. This could be formally taken into account by fitting a random-effects model [23].
- Although perhaps an attractive presentational device, plotting the progress of individual institutions over time on the funnel plot does not provide an appropriate basis for sequential testing due to repeated testing. Methods such as risk-adjusted CUSUMS [24] provide a powerful and formal basis for such sequential monitoring.

On balance we would strongly recommend the use of funnel plots as a useful tool for communicating outcome comparisons.

APPENDIX A

A.1. Proportions

A.1.1. Cross-sectional data. Suppose in each institution that r events are observed out of a sample size of n :

1. The indicator is the observed proportion $y = r/n$ (although both the observed y 's and the control limits may be plotted as percentages).
2. The target proportion θ_0 is such that $\mathbb{V}(Y|\theta_0) = \theta_0(1 - \theta_0)/n$.
3. An interpretable measure ρ of the precision is the sample size, so that $\rho = n$ and $g(\theta_0) = \theta_0(1 - \theta_0)$ in (1).
4. Exact limits can be obtained from the inverse Binomial distribution $F^{-1}(p|\theta_0, n)$, usually defined as the smallest integer r_p such that $P(R \leq r_p) = F(r_p|\theta_0, n) > p$. This means that $F(r_p - 1|\theta_0, n) < p$, and so there exists an α such that $\alpha F(r_p - 1|\theta_0, n) + (1 - \alpha)F(r_p|\theta_0, n) = p$, i.e.

$$\alpha = \frac{F(r_p|\theta_0, n) - p}{f(r_p|\theta_0, n)}$$

where $f(r_p|\theta_0, n) = F(r_p|\theta_0, n) - F(r_p - 1|\theta_0, n)$. It is then natural to use as interpolated control limits a weighted average $y_p(\theta_0, n) = (\alpha(r_p - 1) + (1 - \alpha)r_p)/n = (r_p - \alpha)/n$, clearly showing the role of the continuity adjustment α . Functions to calculate these quantities should be available in standard software: approximate normal limits follow immediately from (3).

A.1.2. Change in proportions. Suppose we have two measures for each institution: r_1, n_1 in a baseline period and r_2, n_2 in a subsequent period, and we wish to assess the change in the underlying proportion from π_1 to π_2 . Three different measures might be of interest: the difference in proportions, the ratio of proportions, or the odds ratio. Normal approximations are used throughout, and for low (especially zero) counts one might add 0.5 to all r 's and 1 to all n 's in order to stabilize the estimates.

Difference in proportions:

1. The indicator is the observed difference in proportion $y = r_2/n_2 - r_1/n_1$.
2. The target θ_0 is for the difference $\pi_2 - \pi_1$. Since $\mathbb{V}(Y|\pi_1, \pi_2) = \pi_2(1 - \pi_2)/n_2 + \pi_1(1 - \pi_1)/n_1$, if we reparameterize in terms of the mean proportion $\pi_m = (\pi_2 + \pi_1)/2$, so that $\pi_2 = \pi_m + \theta_0/2, \pi_1 = \pi_m - \theta_0/2$, we obtain

$$\mathbb{V}(Y|\theta_0, \pi_m) = \frac{(\pi_m + \theta_0/2)(1 - \pi_m - \theta_0/2)}{n_2} + \frac{(\pi_m - \theta_0/2)(1 - \pi_m + \theta_0/2)}{n_1} \tag{A1}$$

There is a nuisance parameter π_m that strictly speaking should be estimated conditional on the target $\pi_2 - \pi_1 = \theta_0$, but no problems should arise from using a standard estimator $\hat{\pi}_m = (r_1 + r_2)/(n_1 + n_2)$ for π_m in (A1).

3. An interpretable measure ρ of the precision can be obtained as follows. Suppose an institution had a fixed sample size $n_1 = n_2 = N$ per period, and observed proportions that were both equal to the overall mean proportion in the entire sample, estimated by $\hat{\pi}_m = \sum(r_{i1} + r_{i2})/\sum(n_{i1} + n_{i2})$. Then this institution would have $\mathbb{V}(Y|\theta_0, \hat{\pi}_m) = g(\theta_0)/N$, where

$$g(\theta_0) = [(\hat{\pi}_m + \theta_0/2)(1 - \hat{\pi}_m - \theta_0/2) + (\hat{\pi}_m - \theta_0/2)(1 - \hat{\pi}_m + \theta_0/2)];$$

for targets near 0, $g(\theta_0) \approx 2\hat{\pi}_m(1 - \hat{\pi}_m)$. Therefore if we take $\rho = g(\theta_0)/\text{Var}(Y|\theta_0, \hat{\pi}_m)$ we can interpret ρ as, approximately, the sample size per period.

4. Approximate normal limits follow immediately from (3).

Ratio in proportions:

1. The indicator is the observed ratio in proportions $y = (r_2/n_2)/(r_1/n_1)$.
2. The target θ_0 is for the ratio π_2/π_1 , although it is convenient to work on a logarithmic scale so that $\log \theta_0$ is a target for $\log(Y)$. Since $\mathbb{V}(\log Y|\pi_1, \pi_2) \approx (1 - \pi_2)/(n_2\pi_2) + (1 - \pi_1)/(n_1\pi_1)$, if we reparameterize in terms of the geometric mean proportion $\pi_g = \sqrt{\pi_2\pi_1}$, so that $\pi_2 = \pi_g\theta_0^{1/2}$, $\pi_1 = \pi_g\theta_0^{-1/2}$, we obtain

$$\mathbb{V}(\log Y|\theta_0, \pi_g) = \frac{\theta_0^{-1/2} - \pi_g}{n_2\pi_g} + \frac{\theta_0^{1/2} - \pi_g}{n_1\pi_g} \tag{A2}$$

The nuisance parameter π_g can be estimated by $\hat{\pi}_g = \sqrt{(r_1r_2)/(n_1n_2)}$, which can then be plugged into (A2).

3. For an interpretable measure of ρ , suppose as before that an institution had a fixed sample size $n_1 = n_2 = N$ per period, and observed proportions that were both equal to the overall geometric mean proportion in the entire sample, say estimated by $\hat{\pi}_g = \sqrt{(\sum r_{i1} \sum r_{i2})/(\sum n_{i1} \sum n_{i2})}$. Then this institution would have $\mathbb{V}(\log Y|\theta_0, \hat{\pi}_g) = g(\theta_0)/N$, where

$$g(\theta_0) = \frac{\theta_0^{-1/2} + \theta_0^{1/2} - 2\hat{\pi}_g}{\hat{\pi}_g};$$

for targets near 1, $g(\theta_0) \approx 2(1 - \hat{\pi}_g)/\hat{\pi}_g$. Therefore if we take $\rho = g(\theta_0)/\mathbb{V}(\log Y|\theta_0, \hat{\pi}_g)$ we can again interpret ρ as, approximately, the sample size per period.

4. Approximate normal limits follow immediately from (3): these will be on a logarithmic scale and so need to be plotted appropriately.

Odds ratios of proportions:

1. The indicator is the observed odds ratio in proportions $y = (r_2/(n_2 - r_2))/(r_1/(n_1 - r_1))$.
2. The target θ_0 is for the odds ratio $(\pi_2/(1 - \pi_2))/(\pi_1/(1 - \pi_1))$, although it is convenient to work on a logarithmic scale so that $\log \theta_0$ is a target for $\log Y$. A standard result gives us $\mathbb{V}(\log Y) \approx 1/r_2 + 1/(n_2 - r_2) + 1/r_1 + 1/(n_1 - r_1)$.

3. For an interpretable measure of ρ , suppose an institution had a fixed sample size $n_1 = n_2 = N$ per period, and observed proportions that were both equal to the overall mean proportion in the entire sample $\hat{\pi}_m$, as defined previously. Then this institution would have $\mathbb{V}(\log Y) \approx g(\theta_0)/N$, where

$$g(\theta_0) = \frac{2}{\hat{\pi}_m(1 - \hat{\pi}_m)}$$

Therefore if we take $\rho = g(\theta_0)/\mathbb{V}(\log Y)$ we can interpret ρ as, approximately, the sample size per period.

4. Approximate normal limits follow immediately from (3): these will be on a logarithmic scale and so need to be plotted appropriately.

A.2. Standardized rates

Two types of standardized rate data are considered: indirectly standardized data which leads to a standardized event ratio such as a standardized mortality ratio (SMR), and directly standardized data that leads to a rate per, say, 1000 individuals. The main exposition is in terms of indirectly standardized rates, and at the end of this section it is shown how to adapt the methods to directly standardized rates.

A.2.1. Cross-sectional data. Suppose in each institution that O events are observed and that indirect standardization specifies an expectation E .

1. The indicator is the observed SMR, $y = O/E$ (although both the observed y 's and the control limits might be plotted as $100 \times \text{SMR}$).
2. The target SMR, θ_0 , implies that in-control institutions will have $O \sim \text{Poisson}(\theta_0 E)$, and hence is such that $\mathbb{V}(Y|\theta_0) = \theta_0/E$.
3. An interpretable measure ρ of the precision is the expectation, so that $\rho = E$ and $g(\theta_0) = \theta_0$ in (1).
4. Exact limits can be obtained using the inverse Poisson($\theta_0 E$) distribution function $o_p = F^{-1}(p|\theta_0 E)$, adopting the interpolation procedure defined in Appendix A.1.1, and setting the control limits as $y_p(\theta_0, \rho) = (o_p - \alpha)/E$. Approximate normal limits follow immediately from (3).

A.2.2. Change in standardized rates. Suppose we have two measures for each institution: O_1, E_1 in a baseline period and O_2, E_2 in a subsequent period, and we wish to assess the change in the underlying SMR. We shall only consider the ratio of rates: exact methods based on a conditional argument are available if $E_1 = E_2$, and otherwise normal approximations are used, in which case for low (especially zero) counts one might add 0.5 to all O 's and E 's.

Ratio of SMRs, exact results assuming $E_1 = E_2 = E$:

1. The indicator is the observed ratio in SMRs $y = (O_2/E_2)/(O_1/E_1) = O_2/O_1$.
2. The target θ_0 is for the true SMR ratio denoted λ_2/λ_1 .
3. For reasons explained below, the precision ρ can be taken as the average observed count $(O_1 + O_2)/2$.

4. For in-control institutions we have $O_1 \sim \text{Poisson}(\lambda_1 E)$, $O_2 \sim \text{Poisson}(\theta_0 \lambda_1 E)$. From a standard result in conditional inference [25, p. 93] the distribution of O_2 conditional on $O_1 + O_2$ is Binomial with underlying proportion $\theta_0/(1 + \theta_0)$. Hence we can use the methods of Section A.1.1 to obtain critical limits $y_p(\theta_0/(1 + \theta_0), O_1 + O_2)$ for the observed proportion $r = O_2/(O_1 + O_2)$; it is then natural to take $\rho = (O_1 + O_2)/2$, and transform back to the original rate scale via $y = r/(1 - r)$.

Ratio of SMRs, general normal approximations:

1. The indicator is the observed ratio in SMRs $y = (O_2/E_2)/(O_1/E_1)$.
2. The target θ_0 is for the underlying rate ratio λ_2/λ_1 , although it is convenient to work on a logarithmic scale so that $\log \theta_0$ is a target for $\log Y$. Since $\mathbb{V}(\log Y|\lambda_1, \lambda_2) \approx 1/(E_2 \lambda_2) + 1/(E_1 \lambda_1)$, if we reparameterize in terms of the geometric mean SMR $\lambda_g = \sqrt{\lambda_2 \lambda_1}$, so that $\lambda_2 = \lambda_g \theta_0^{1/2}$, $\lambda_1 = \lambda_g \theta_0^{-1/2}$, we obtain

$$\mathbb{V}(\log Y|\theta_0, \lambda_g) = \frac{\theta_0^{-1/2}}{E_2 \lambda_g} + \frac{\theta_0^{1/2}}{E_1 \lambda_g} \quad (\text{A3})$$

In this case a maximum likelihood estimator for the nuisance parameter λ_g is easily obtainable under the null hypothesis that $\lambda_2/\lambda_1 = \theta_0$: $\hat{\lambda}_g = (O_1 + O_2)/(\theta_0^{1/2} E_2 + \theta_0^{-1/2} E_1)$, which can then be plugged into (A3).

3. For an interpretable measure of ρ , suppose as before that an institution had a fixed expectation $E_1 = E_2 = E$ per period, and whose geometric mean SMR was equal to the overall mean in the entire sample, say estimated by $\hat{\lambda}_g = (\sum O_{i1} + \sum O_{i2})/(\sum E_{i1} + \sum E_{i2})$. Then this institution would have $\text{Var}(\log Y|\theta_0, \hat{\lambda}_g) = g(\theta_0)/E$, where

$$g(\theta_0) = (\theta_0^{-1/2} + \theta_0^{1/2})/\hat{\lambda}_g$$

Therefore if we take $\rho = g(\theta_0)/\mathbb{V}(\log Y|\theta_0, \hat{\lambda}_g)$ we can interpret ρ as, approximately, the expectation per period.

4. Approximate normal limits follow immediately from (3): these will be on a logarithmic scale and so need to be plotted appropriately.

A.2.3. Directly standardized rates. We assume these are reported as a rate per (say) 1000 individuals, and that the rate has been transformed to a proportion y between 0 and 1. The measure of the associated error may be reported in a variety of ways, and possibilities include:

1. A confidence interval (c_1, c_2) for the rate.
2. The size of population N .
3. The observed number of events O upon which the rate is based.

Given any one of these the others can (at least approximately) be obtained so the same plotting procedure can, in theory, be used in all situations. For example, given a symmetric confidence interval we can interpret its width $(c_2 - c_1)$ as being 2×1.96 standard errors, where the standard error is given by $\sqrt{y(1-y)/N}$, and hence the effective population N obtained.

Funnel plots might then be based on those for proportions (Section A.1) using y and N . Alternatively, for rare events we might assume $O \sim \text{Poisson}(\lambda N)$ for some true rate λ , we

may use those for indirectly standardized rates by taking N as E in the methods for indirectly standardized rates given above.

A.3. Continuous responses

A.3.1. Cross-sectional data. Suppose in each institution that a measure is observed on a sample of n individuals, leading to a sample mean y with standard error s .

1. The indicator is the observed mean y .
2. The target proportion θ_0 is such that $\mathbb{V}(Y|\theta_0) = s^2$.
3. If an estimated common standard deviation $\hat{\sigma}$ is available, then s can be set equal to $\hat{\sigma}/\sqrt{n}$ and ρ can be taken as the effective sample size n . If there is no common standard deviation then one could take $\hat{\sigma}^2 = \sum n_i s_i^2 / I$, and set ρ as the effective sample size $n^{\text{eff}} = (\hat{\sigma}/s)^2$. Thus $g(\theta_0) = \hat{\sigma}^2$ in (1).
4. Exact control limits are obtained by setting $y_p = \theta_0 + z_p \sqrt{\hat{\sigma}/\rho}$.

A.3.2. Difference in means. Suppose we have two measures for each institution: y_1, s_1 in a baseline period and y_2, s_2 in a subsequent period, and we wish to assess the change between the underlying means.

1. The indicator is the observed difference in means $y = y_2 - y_1$.
2. The target θ_0 is for the difference in means, and

$$\mathbb{V}(Y|\theta_0) = s_1^2 + s_2^2 \quad (\text{A4})$$

3. An interpretable measure ρ of the precision can be obtained as follows. Suppose an institution had a fixed sample size $n_1 = n_2 = N$ per period, then $\mathbb{V}(Y|\theta_0) = g(\theta_0)/N$, where

$$g(\theta_0) = 2\hat{\sigma}^2$$

Therefore if we take $\rho = g(\theta_0)/\mathbb{V}(Y|\theta_0)$ we can interpret ρ as, approximately, the sample size per period.

4. Exact normal limits follow immediately from (3).

ACKNOWLEDGEMENTS

I am indebted to Martin Bardsley's Screening and Surveillance team at the Commission for Health Improvement for motivating and encouraging this work, Julian Flowers of the Eastern Region Public Health Observatory for providing the teenage pregnancy data, and to Vern Farewell and Ken Rice for helpful comments.

REFERENCES

1. *NHS Performance Indicators*. The Stationery Office: London, 2002. <http://www.performance.doh.gov.uk/nhsperformanceindicators/2002/trust.html> (accessed February, 2004).
2. Goldstein H, Spiegelhalter DJ. Statistical aspects of institutional performance: league tables and their limitations (with discussion). *Journal of the Royal Statistical Society, Series A* 1996; **159**:385–444.
3. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *British Medical Journal* 1998; **317**:1701–1704.
4. Mohammed MA, Cheng KK, Rouse A, Marshall T. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet* 2001; **357**:463–467.

5. Shewhart WA. The application of statistics as an aid in maintaining quality of a manufactured product. *Journal American Statistical Association* 1925; **20**:546–548.
6. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997; **315**:629–634.
7. Stark J, Gallivan S, Lovegrove J, Hamilton JRL, Monro JL, Pollock JCS, Watterson KG. Mortality rates after surgery for congenital heart defects in children and surgeons' performance. *Lancet* 2000; **355**:1004–1007.
8. Stark JF, Gallivan S, Davis K, Hamilton JRL, Monro JL, Pollock JCS, Watterson KG. Assessment of mortality rates for congenital heart defects and surgeons' performance. *Annals of Thoracic Surgery* 2001; **72**:169–174.
9. Spiegelhalter DJ. An investigation into the relationship between mortality and volume of cases: an example in paediatric cardiac surgery between 1991 to 1995. *British Medical Journal* 2002; **324**:261–263.
10. Spiegelhalter DJ. Funnel plots for institutional comparisons (letter). *Quality Safety in Health Care* 2002; **11**: 390–391.
11. Tekkis PP, McCulloch P, Steger AC, Benjamin IS, Poloniecki JD. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. *British Medical Journal* 2003; **326**:786–788.
12. Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology* 2001; **54**:1046–1055.
13. *Coronary Artery Bypass Surgery in New York State, 1997–9*. New York State Department of Health: Albany, NY, 2002. http://www.health.state.ny.us/nysdoh/heart/heart_disease.htm (accessed February, 2004).
14. *The NHS Plan*. The Stationery Office: London, 2000.
15. Spiegelhalter DJ, Aylin P, Evans SJW, Murray GD, Best NG. Commissioned analysis of surgical performance using routine data: lessons from the Bristol Inquiry (with discussion). *Journal of the Royal Statistical Society, Series A* 2002; **165**:191–232.
16. Irwig L, Macaskill P, Berry G, Glasziou P. Bias in meta-analysis detected by a simple, graphical test—graphical test is itself biased. *British Medical Journal* 1998; **316**:470.
17. Wright DE, Sanders H. Biased assessment of bias in meta-analysis, 2003, submitted.
18. Marshall C, Best N, Bottle A, Aylin P. Statistical issues in the prospective monitoring of health outcomes across multiple units. *Journal of the Royal Statistical Society, Series A* 2004; **167**:541–560.
19. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
20. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
21. Healthcare Commission. *Clinical Indicator AS403: Emergency readmissions to hospital within 28 days of discharge, as a percentage of live discharges for patients aged 16 years and over*. London, 2004. <http://ratings.healthcarecommission.org.uk/Indicators.2004/downloads/1403c.pdf> (accessed June, 2004).
22. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B—Statistical Methodology* 2002; **64**:479–498.
23. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* 1997; **127**:764–768.
24. Grigg O, Farewell VT, Spiegelhalter DJ. The use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Statistical Methods in Medical Research* 2003; **12**:147–170.
25. Breslow NE, Day NE. *Statistical Methods on Cancer Research, Volume 2: The Design and Analysis of Cohort Studies*. International Agency for Cancer Research: Lyon, 1987.