The Identification of "Unusual" Health-Care Providers From a Hierarchical Model

Hayley E. JONES and David J. SPIEGELHALTER

It has become common to adopt a hierarchical model structure when comparing the performance of multiple health-care providers. This structure allows some variation in such measures, beyond that explained by sampling variation, to be "normal," in recognition of the fact that risk-adjustment is never perfect. The shrinkage estimates arising from such a model structure also have appealing properties.

It is not immediately clear, however, how "unusual" providers, that is, any with particularly high or low rates, can be identified based on such a model. Given that some variation in underlying rates is assumed to be the norm, we argue that it is not generally appropriate to identify a provider as interesting based only on evidence of it lying above or below the population mean. We note with concern, however, that this practice is not uncommon.

We examine in detail three possible strategies for identifying unusual providers, carefully distinguishing between statistical "outliers" and "extremes." A two-level normal model is used for mathematical simplicity, but we note that much of the discussion also applies to alternative data structures. Further, we emphasize throughout that each approach can be viewed as resulting from a Bayesian or a classical perspective. Three worked examples provide additional insight.

KEY WORDS: Outliers; Posterior tail areas; Provider profiling; Unusual performance.

1. INTRODUCTION

Since the mid-1990s the case for hierarchical models in monitoring the performance of multiple health-care providers, or "provider profiling," has been very strongly argued (Thomas, Longford, and Rolph 1994; Goldstein and Spiegelhalter 1996; Morris and Christiansen 1996; Normand, Glickman, and Gatsonis 1997; Burgess et al. 2000). A hierarchical or multilevel model structure arises from assuming there exist provider-specific "random effects" which are assumed drawn from some common distribution. Such a model is particularly appealing because of the well-documented benefits of the resulting shrinkage estimates of each unit-specific rate (Greenland and Robins 1991; Louis 1991), which improve estimation performance by "borrowing" information from other providers. Theoretical arguments (e.g., James and Stein 1961) and empirical studies (Efron and Morris 1975; Rubin 1980; Tomberlin 1988) have both been used to demonstrate the improved average point predictive ability of shrinkage estimates over crude observed rates; the automatic shrinkage of the estimates toward the overall average means the predictions can be thought of as "adjusting for regression-to-the-mean" (Burgess et al. 2000). Some precision is also gained by the pooling of information, allowing small counts to be dealt with more effectively. Use of a hierarchical model can also be motivated by a desire to account for observed overdispersion, allowing some variation in estimated performance to be "normal" in recognition of imperfect risk-adjustment (Spiegelhalter 2005b). It has been shown, further, that the power of a test for individual recent changes can be improved slightly by shrinkage of each baseline observation (Jones and Spiegelhalter 2009).

Although the advantages of a hierarchical modeling structure may now seem clear, there remains confusion in the literature about how individual units of interest should be identified based on such a model. We note with some concern that several authors have based the identification of "unusual" performance on the posterior probability of a measure being greater (or less) than the population mean, or equivalent measures (Simpson et al. 2003; Smits et al. 2003; Darlow et al. 2005; Racz and Sedransk 2010). We argue that this is not generally appropriate: after making the assumption that some variation around the population mean is the norm, it does not seem reasonable to then use "above or below the average" as the criterion for highlighting providers of interest. In particular, we stress that this procedure does not identify statistical "outliers" as has been suggested (Smits et al. 2003; Racz and Sedransk 2010).

In this article we examine in detail how health-care providers with unusually high rates of some measure (potentially inferior performance, in the context of our examples) might be identified, with reference to a simple and commonly used hierarchical model for cross-sectional data. The strategies discussed are of course easily generalized to identify potentially superior performance, which might also be of interest. A careful distinction is drawn between statistical *outliers* and *extremes* in Section 2. While an "extreme" observation might well be accommodated within the assumed model, for a provider to be deemed "outlying" it essentially must lie beyond the range allowed by the model. Relevant literature in which authors have attempted to identify unusual performance based on hierarchical models is then briefly reviewed.

Hayley E. Jones is Research Fellow, School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol, BS8 2PS, U.K. (E-mail: *hayley.jones@bristol.ac.uk*). David J. Spiegelhalter is Senior Scientist, MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge, CB2 0SR, U.K., and Winton Professor of the Public Understanding of Risk, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, U.K. This work was funded by an MRC Ph.D. studentship at the MRC Biostatistics Unit, Cambridge, U.K.

A simple two-level normal model is described in Section 3.1. With reference to this model, three alternative approaches to identifying potentially poor providers are outlined in detail in Sections 3.2–3.4. Throughout Section 3 we emphasize and demonstrate that each of the three approaches outlined can equally well be viewed as resulting from a Bayesian or classical multilevel modeling framework. A more general discussion of multilevel models from a classical perspective is provided by Goldstein (2003).

These three approaches to identifying unusual performance are each demonstrated using worked examples in Section 4. We consider teenage conception rates in English Local Authorities, rates of *Clostridium difficile* in NHS Trusts, and, finally, mortality rates following heart surgery in New York State hospitals. As we will discuss, these examples differ with respect to a key parameter ρ, a form of intraclass correlation coefficient that measures the amount of "true" variability in underlying rates relative to total observed variability. Funnel plots (Spiegelhalter 2005a; Schulman, Spiegelhalter, and Parry 2008) are used to illustrate the properties of each approach under the different estimated values of ρ. Concluding remarks are then made in Section 5.

2. APPROACHES TO IDENTIFYING UNUSUAL PERFORMANCE

We distinguish between three approaches to identifying unusual performance using cross-sectional data, namely:

Approach 1: *Identify outliers to the common mean model.*
Approach 2: *Identify outliers to the random effects distribution.*
Approach 3: *Identify extremes in the random effects distribution.*

Approach 1 involves using a common mean (i.e., nonhierarchical) model as the null hypothesis, and simply testing for deviations from this. Further details are provided, for example, by Jones, Ohlssen, and Spiegelhalter (2008), who used this simple method to demonstrate how the false discovery rate (FDR) (Benjamini and Hochberg 1995) can be controlled to account for multiple testing in this context.

As discussed in the Introduction, there are strong arguments in favor of assuming instead a random effects model, which focuses attention on approaches 2 and 3. Approach 2 takes the random effects distribution as a null model and identifies deviations from this, while approach 3 is based on posterior tail areas of the individual random effects. We will argue that either approach is reasonable, but that analysts must be very clear about which they are using and why. Essentially, the choice should be based on the selected modeling strategy, as discussed by Ohlssen, Sharples, and Spiegelhalter (2007):

Hypothesis Testing Strategy. The random effects distribution is posited as a null model, allowing for some variation around the population mean to be "normal," due to unknown factors beyond the providers' control (i.e., imperfect risk-adjustment). Some true outliers to this distribution may well still exist, so that approach 2 is sensible. This approach is used, for example, by the U.K. Care Quality Commission (2009) as part of the "Annual Health Check."

Estimation Strategy. The hierarchical model is constructed to be encompassing, such that all providers ("unusual" or not) are accommodated. It does not then make sense to try to identify outliers except for model checking purposes, since by construction there should be none. Approach 3 is then appropriate.

When using approach 3 the crucial element is what is considered "extreme." Some authors have used methods equivalent to identifying a provider as unusual based on a large posterior probability of its rate being greater (or less) than average (Simpson et al. 2003; Smits et al. 2003; Darlow et al. 2005; Racz and Sedransk 2010). The crucial point is that such providers need not be "extremes"; in fact, if a provider is large enough, then its posterior interval is unlikely to include the population mean.

In contrast, Morris and Christiansen (1996) worked within the estimation framework but suggested that comparisons with the population mean are inappropriate, arguing instead in favor of comparisons with an external target or interval standard (Morris and Christiansen 1996; Christiansen and Morris 1997b; Burgess et al. 2000). Normand, Glickman, and Gatsonis (1997) made similar recommendations, although suggested that, in the absence of an external target, indices such as the posterior probability of each rate being greater than 1.5 times the median might be used. Similarly, in a multivariate setting in which inferences are made about a latent performance variable based on multiple observed measures, profiling has been based on the posterior probability of this latent variable lying in, for example, the top 10 or 20 percent (Landrum, Normand, and Rosenheck 2003; Teixeira-Pinto and Normand 2008). We concur with such recommendations if an estimation modeling strategy is employed, but suggest that the classical interpretation of these posterior tail areas is also in need of consideration.

When random effects modeling has been adopted in the literature, it is sometimes not clear which approach has been used, due to authors being nonspecific about which standard errors were used to construct prediction limits or confidence intervals. For example, Racz and Sedransk (2010) and Simpson et al. (2003) each referred to the "standard errors" of their shrinkage estimators, while in fact this is an ambiguous term in this context. We will show later that the choice of standard error corresponds to choosing between approach 2 and the particular instance of approach 3 whereby "above or below the average" is used to define extreme performance.

3. THEORY FOR A SIMPLE TWO-LEVEL NORMAL MODEL

3.1 Notation

Before considering each of the three approaches in some detail, we first introduce notation, and details of the simple two-level normal model which is often used in this context.

Assume a performance measure Y_i is observed on each of $i = 1, \ldots, m$ health-care providers, and that

$$Y_i|\theta_i \overset{\text{indep}}{\sim} \text{Normal}(\theta_i, \sigma_i^2), \qquad \theta_i \overset{\text{indep}}{\sim} \text{Normal}(\mu, \tau^2), \quad (1)$$

where the σ_i's are assumed known. This is often referred to as an unbalanced one-way analysis of variance (ANOVA) model.

From Bayes's rule, it can easily be shown that

$$\theta_i | y_i, \mu, \tau \sim \text{Normal}(\hat{\theta}_i, w_i \sigma_i^2), \tag{2}$$

where

$$\hat{\theta}_i = w_i y_i + (1 - w_i)\mu \tag{3}$$

and

$$w_i = \frac{\tau^2}{\sigma_i^2 + \tau^2} \tag{4}$$

can be seen to be a classical intraclass correlation coefficient.

$\hat{\theta}_i$ is called a shrinkage estimate of θ_i, since the observed y_i is "shrunk" in toward the population mean, μ, with weight $0 \le w_i \le 1$ determining the degree of shrinkage. For $\tau = 0$ (identical rates in all providers), $w_i = 0$ so that $\hat{\theta}_i$ is appropriately set to the population mean. As τ increases to infinity, w_i tends to 1 so that there is no pooling of information and $\hat{\theta}_i$ is simply set to y_i. The amount of shrinkage in each individual provider is also dependent on its "size," with greatest shrinkage toward μ when σ_i^2 is large (i.e., for a "small" provider).

Adopting an empirical Bayes (EB) approach, like many other authors in the performance monitoring literature (Greenland and Robins 1991; Howley and Gibberd 2003), the unknown μ and τ^2 are estimated from the marginal distribution of the data

$$Y_i \overset{\text{indep}}{\sim} \text{Normal}(\mu, \sigma_i^2 + \tau^2)$$

and plugged into (2)–(4) as if known. In our experience, these parameters can usually be estimated quite precisely in a performance monitoring context (in contrast to the area of meta-analysis, where the number of contributing data points is generally much smaller) so that ignoring uncertainty in their estimation will have little influence on the identification of unusual providers. However, in the case of only a small number of providers m, adjustment of EB estimates to account for uncertainty should be considered (Carlin and Louis 2000) or a fully Bayesian approach used.

Denoting $a_i \equiv 1/\sigma_i^2$, the well-known DerSimonian and Laird estimate of τ^2 is

$$\hat{\tau}^2 = \max\left\{0, \frac{\sum a_i (y_i - \bar{y}_w)^2 - (m-1)}{\sum a_i - \sum a_i^2 / \sum a_i}\right\}, \tag{5}$$

where

$$\bar{y}_w = \frac{\sum a_i y_i}{\sum a_i},$$

while the population mean can be estimated by

$$\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i},$$

a weighted average of the y_i's.

If using approach 2, it is desirable in practice to estimate τ using robust methods, so that truly divergent providers are not accommodated by the null model (Spiegelhalter 2005b; Ohlssen, Sharples, and Spiegelhalter 2007). However, for ease of exposition in comparing different profiling strategies we will simply use (5) in our worked examples.

3.2 Approach 1: *Identify Outliers to the Common Mean Model*

The common mean model is assumed here, that is, τ is assumed equal to 0.

3.2.1 Bayesian Perspective

We adopt the Bayesian approach to model checking outlined by Gelman and Hill (2007), in which a Bayesian p-value is calculated by comparing the observed statistic with its predictive distribution obtained from integrating out unknown parameters. If we assume μ is essentially known, the predictive distribution of each Y_i is simply $Y_i^{\text{pred}} \sim \text{Normal}(\mu, \sigma_i^2)$. The predictive p-value assesses the plausibility under this distribution of seeing a rate as high as that observed. It is equal to

$$\begin{aligned} p_i^{(1)} &\equiv P(Y_i^{\text{pred}} > y_i) \\ &= 1 - \Phi\left(\frac{y_i - \mu}{\sigma_i}\right) \\ &= \Phi\left(\frac{\mu - y_i}{\sigma_i}\right). \end{aligned} \tag{6}$$

3.2.2 Classical Perspective

We test whether the underlying performance θ_i in provider i is equal to the population mean. Specifically, our model is $Y_i \sim \text{Normal}(\theta_i, \sigma_i^2)$ and we test $H_0 : \theta_i = \mu$ against the one-sided alternative $H_1 : \theta_i > \mu$ for each provider, $i = 1, \ldots, m$.

The test statistic to be referred to the standard normal distribution is simply

$$Z_i^{(1)} = \frac{y_i - \mu}{\sigma_i},$$

so that the p-value of interest $= 1 - \Phi(Z_i^{(1)}) = p_i^{(1)}$ as in (6).

3.3 Approach 2: *Identify Outliers to the Random Effects Distribution*

3.3.1 Bayesian Perspective

Under the random effects model, provider i's predictive distribution is given by

$$Y_i^{\text{pred}} \sim \text{Normal}(\mu, \sigma_i^2 + \tau^2).$$

The predictive p-value used to identify rates higher than plausible under this null distribution is therefore

$$\begin{aligned} p_i^{(2)} &= \Phi\left(\frac{\mu - y_i}{\sqrt{\sigma_i^2 + \tau^2}}\right) \\ &= \Phi\left(\sqrt{1 - w_i}\left(\frac{\mu - y_i}{\sigma_i}\right)\right). \end{aligned} \tag{7}$$

Note that the unit-specific θ_i plays no part here, as it has been integrated out in order to obtain the predictive distribution.

3.3.2 Classical Perspective

Consider the shrinkage estimate $\hat{\theta}_i$ as defined by (3). If θ_i comes from the same distribution as the other providers, then, unconditionally on θ_i, $\hat{\theta}_i$ is an unbiased estimator of the population mean μ. Otherwise, the unconditional expectation of $\hat{\theta}_i$ takes some other value, say $E(\hat{\theta}_i) \equiv \mu_i$.

We treat $\hat{\theta}_i$ as an estimate of μ_i (not of θ_i). It is used to test $H_0 : \mu_i = \mu$ against the alternative $H_1 : \mu_i > \mu$, in each provider. If H_0 is rejected in favor of H_1 for provider i, then it is concluded that θ_i appears to be too large to have come from the null random effects distribution.

The predictive distribution of $\hat{\theta}_i$ under H_0 is

$$\hat{\theta}_i \sim \text{Normal}\big(\mu, V_{Y_i,\theta_i}(\hat{\theta}_i)\big),$$

where $V_{Y_i,\theta_i}(\hat{\theta}_i)$ is the variance of $\hat{\theta}_i$ over the joint distribution of Y_i and θ_i given H_0, and is given by

$$\begin{aligned} V_{Y_i,\theta_i}(\hat{\theta}_i) &= w_i^2 V_{Y_i,\theta_i}(Y_i) \\ &= \frac{\tau^4}{\sigma_i^2 + \tau^2} \\ &= w_i \tau^2. \end{aligned} \tag{8}$$

The classical test statistic is therefore

$$\begin{aligned} Z_i^{(2)} &= \frac{\hat{\theta}_i - \mu}{\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i)}} \\ &= \frac{y_i - \mu}{\sqrt{\sigma_i^2 + \tau^2}}, \end{aligned}$$

resulting in the p-value $1 - \Phi(Z_i^{(2)}) = p_i^{(2)}$, exactly as in the Bayesian result (7).

Note that it is the case in general, regardless of the distribution of Y_i, that

$$\frac{\hat{\theta}_i - \mu}{\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i)}} = \frac{y_i - \mu}{\sqrt{V_{Y_i,\theta_i}(Y_i)}} \tag{9}$$

for any shrinkage estimator of the form $\hat{\theta}_i = \kappa_i Y_i + (1 - \kappa_i)\mu$ with weight κ_i, since $V_{Y_i,\theta_i}(\hat{\theta}_i) = \kappa_i^2 V_{Y_i,\theta_i}(Y_i)$ and $\hat{\theta}_i - \mu = \kappa_i(Y_i - \mu)$. In particular, this holds for a two-level Poisson model with $O_i | r_i \overset{\text{indep}}{\sim} \text{Poisson}(r_i E_i)$ for known E_i, where the r_i's are drawn from a common gamma distribution. In this context, we define $Y_i \equiv O_i / E_i$ and $\hat{\theta}_i \equiv E(r_i | O_i)$. Poisson–gamma EB models have been used, for example, by Simpson et al. (2003) in performance monitoring and are also popular in the small area estimation literature (McPherson et al. 1982; Clayton and Kaldor 1987; Coory and Gibberd 1998). Fully or approximate Bayesian versions of this model are also commonly fitted in performance monitoring (Goldstein and Spiegelhalter 1996; Christiansen and Morris 1997a; Normand, Glickman, and Gatsonis 1997).

3.4 Approach 3: *Identify Extremes in the Random Effects Distribution*

3.4.1 Bayesian Perspective

It is natural to use the posterior distribution of θ_i, as given by (2), to identify extremes of the distribution. A provider might be considered "extreme" if its rate has a large posterior probability of being greater than some external target or specified quantile of the random effects distribution, say t. This posterior probability is

$$\begin{aligned} P(\theta_i > t | y_i) &= 1 - \Phi\left(\frac{t - E(\theta_i | y_i)}{\sqrt{V(\theta_i | y_i)}}\right) \\ &= 1 - \Phi\left(\frac{t - \hat{\theta}_i}{\sigma_i \sqrt{w_i}}\right). \end{aligned} \tag{10}$$

As discussed previously, some authors have (we believe inappropriately) essentially used this approach with $t = \mu$, the population mean. For this special case, the following holds:

$$\begin{aligned} P(\theta_i > \mu | y_i) &= 1 - \Phi\left(\frac{\mu - \hat{\theta}_i}{\sigma_i \sqrt{w_i}}\right) \\ &= 1 - \Phi\left(\frac{\mu - y_i}{\sqrt{\frac{\sigma_i^2}{\tau^2}(\sigma_i^2 + \tau^2)}}\right) \\ &= 1 - \Phi\left(\sqrt{w_i}\left(\frac{\mu - y_i}{\sigma_i}\right)\right) \\ &\equiv 1 - p_i^{(3)}, \end{aligned} \tag{11}$$

say, although we note that $p_i^{(3)}$ is not a p-value in this context, but rather a Bayesian posterior tail area. Following this approach, provider i is classified as potentially extreme if $p_i^{(3)}$ is small.

3.4.2 Classical Perspective

If $\hat{\theta}_i$ is treated as an estimate of θ_i rather than of the *unconditional* mean μ_i, when constructing a confidence interval or using it as a basis for a hypothesis test we should take account of the fact that it is biased by using the appropriate error measure. As we will discuss below, this is not $\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i)}$, but rather $\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i)}$, or equivalently (since when averaging over the distribution of the true parameters, a shrinkage estimator of the population mean is unbiased) the root mean squared error $\sqrt{E_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i)^2}$. This error term appropriately measures how far we expect $\hat{\theta}_i$ to be from what it is trying to estimate in this case.

As such, if we wish to test $H_0 : \theta_i < t$ versus the alternative $H_1 : \theta_i > t$, then the appropriate test statistic is

$$Z_i^{(3)} = \frac{\hat{\theta}_i - t}{\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i)}}.$$

Now, the conditional mean squared error for known Y_i can easily be shown to be

$$E_{\theta_i | Y_i}(\hat{\theta}_i - \theta_i)^2 = V(\theta_i | Y_i).$$

This result is in fact well known in decision theory: assuming known hyperparameters, the posterior risk for a squared error loss function is minimized by the posterior mean and this minimum risk is equal to the posterior variance (Lee 1997). It follows that, independently of the normal assumption,

$$\begin{aligned} V_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i) &= E_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i)^2 \\ &= E_{Y_i}\left[E_{\theta_i|Y_i}(\hat{\theta}_i - \theta_i)^2\right] \\ &= E_{Y_i}[V(\theta_i|Y_i)]. \end{aligned} \tag{12}$$

For the one-way ANOVA model, $V(\theta_i|Y_i)$ does not depend on the data Y_i, so that $V_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i)$ is equal to the posterior variance itself, $w_i\sigma_i^2$.

Given this result, the appropriate p-value for this test, $1 - \Phi(Z_i^{(3)})$, is equal to one minus the value defined by (10), or $p_i^{(3)}$ in the particular case of the commonly used threshold $t = \mu$. Thus under certain restrictions the Bayesian posterior tail area can be shown to be a classical p-value for testing a specific null hypothesis, $\theta_i < t$. For the threshold $t = \mu$ this approach simply tests whether the institution is in the top or bottom half of the random effects distribution.

In more general circumstances these classical and Bayesian approaches will not be exactly equivalent, because the posterior variance may depend on the observed statistic. For example, for the Poisson–gamma model,

$$V(r_i|O_i) = \frac{O_i + \mu^2/\tau^2}{(E_i + \mu/\tau^2)^2},$$

clearly depending on the observed count, O_i. The error measure used for the classical test is therefore not quite identical to that which a Bayesian would use in this instance.

3.5 Standard Errors of Random Effects

From a classical perspective, Goldstein (2003) distinguished between two alternative standard errors of random effects:

(1) the unconditional or *diagnostic* standard error,

$$\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i - \mu)} = \sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i)},$$

equal to $\tau\sqrt{w_i}$ for the two-level normal model, as shown in (8).

(2) the conditional or *comparative* standard error,

$$\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i)},$$

which, from (12), is equal to $\sigma_i\sqrt{w_i}$ or equivalently $\tau\sqrt{1 - w_i}$ here.

Goldstein suggested that diagnostic standard errors should be used when examining the distributional properties of residuals, while comparative standard errors are required for making inferences about the true random effects. In agreement with this, level-2 residuals from hierarchical models have commonly been standardized by their diagnostic standard errors when checking the assumption of normality of random effects and detecting outliers (Lange and Ryan 1989; Hardy and Thompson 1998; Langford and Lewis 1998). We reinforce Goldstein's message, by stressing that $\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i)}$ is the appropriate standard error to use for the identification of outliers to the random effects distribution. $\sqrt{V_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i)}$ might, however, be of interest for making inferences about θ_i, such as assessing whether it is "extreme."

From result (12) we see that $V_{Y_i,\theta_i}(\hat{\theta}_i - \theta_i)$ also has a Bayesian interpretation as the expectation of the posterior variance. As such, use of the comparative standard error in the identification of "unusual" performance is roughly (or exactly, for the normal model) equivalent to examining Bayesian posterior tail areas. These tail areas can be used to make inferences about the performance of individual providers and possibly to identify extremes, but do not indicate whether a provider's performance is outlying.

3.6 The Effect of ρ

We have seen (Equations (6), (7), and (11)) that the three different approaches involve the following p-values:

1. *Identify outliers to the common mean model*:

$$p_i^{(1)} = \Phi\left(\frac{\mu - y_i}{\sigma_i}\right).$$

2. *Identify outliers to the random effects distribution*:

$$p_i^{(2)} = \Phi\left(\sqrt{1 - w_i}\left(\frac{\mu - y_i}{\sigma_i}\right)\right).$$

3. *Identify extremes in the random effects distribution* (if inappropriately using "above or below average" as the definition of extreme):

$$p_i^{(3)} = \Phi\left(\sqrt{w_i}\left(\frac{\mu - y_i}{\sigma_i}\right)\right).$$

Clearly, if w_i is small, then $p_i^{(1)}$ will be very similar to $p_i^{(2)}$, while if w_i is large, then it will instead be similar to $p_i^{(3)}$.

Note that $w_i = \mathrm{Corr}(Y_{i1}, Y_{i2})$, the correlation between any two measures made on the same provider, assuming constant known sampling variance σ_i^2. We define a summary measure

$$\rho \equiv \frac{\tau^2}{\tau^2 + \bar{\sigma}^2}, \tag{13}$$

the correlation between two measures made on an "average sized" provider, where $\bar{\sigma}^2$ is the sample mean of the σ_i^2's. The parameter ρ provides a measure of the magnitude of the between-provider variability relative to average total variability. To obtain an estimate of ρ, we will simply plug the estimate $\hat{\tau}^2$ into (13).

Across the entire population of providers, if ρ is close to 0, then approaches 1 and 2 will highlight similar units as being of interest. This will be the case when the quality of risk-adjustment applied before profiling is high. In contrast, since the posterior distribution of each θ_i will be narrow and centered close to μ, the $p_i^{(3)}$'s in this case will tend to be close to 0.5, so that very few units above or below average are identified.

If ρ is close to 1, that is, the amount of between-provider variability is large, then the converse is true. Following approach 2, the null random effects distribution allows for this variability, the $p_i^{(2)}$'s therefore tending to be close to 0.5. However, we can

be relatively sure about which half of the random effects distribution each provider lies in. As a result, the $p_i^{(3)}$'s will hardly "correct" for the overdispersion at all: providers identified as outlying to the common mean model will also tend to be identified as having rates above or below average.

4. EXAMPLES

We now consider three worked examples: rates of teenage conceptions in England in 2004, rates of *Clostridium difficile* in English National Health Service (NHS) Trusts in the period October–December 2006, and mortality rates following coronary artery bypass graft (CABG) surgery in New York State in 2003. In each case, data on multiple healthcare providers are publicly available. It is of interest to identify "unusual" Local Authorities, NHS Trusts, and hospitals, respectively.

In each case we transform observed (O_i) and "expected" (E_i) counts to log relative risks $y_i = \log(O_i/E_i), i = 1, \ldots, m$, which we assume are approximately normally distributed. The σ_i^2's of (1) are assumed known and equal to $1/E_i$. Standardization by the expected count E_i is a standard approach to adjusting for risk factors over which the healthcare providers cannot reasonably be held accountable, such as patient risk at admission or demographics. An alternative "exact" Poisson model might of course be fitted, and is theoretically more appealing since the normal approximation will tend to be poor for small counts. We will return to this briefly in Section 5.

Using particular risk-adjustment models, that is, formulas to calculate the E_i's, these three datasets have quite different estimated ρ's, allowing a demonstration of the effect of this parameter as discussed earlier. Some further information about the three examples, including details of the risk-adjustment employed and a discussion of the effect of ρ on a related type of analysis, can be found in the article by Jones and Spiegelhalter (2009).

In each case, funnel plots (Spiegelhalter 2005a; Schulman, Spiegelhalter, and Parry 2008) are used to illustrate the results of each profiling technique graphically. In these plots, each observed performance measure y_i is plotted against a measure of its precision, E_i. Prediction limits corresponding exactly to particular thresholds for the derived p-values are plotted, forming a funnel shape in recognition of the increased sampling variability expected in smaller units. Using an appropriate p-value threshold, providers lying outside of the funnel shape might be considered interesting for further investigation.

In applying approach 3, we will assume that the threshold t is set to the population mean μ, in order to demonstrate the characteristics of this commonly used procedure. However, we emphasize that we believe it is much more appropriate to use some other value in practice.

4.1 Teenage Conceptions in Local Authorities

Figure 1 shows the results of applying each approach to $m = 352$ observed teenage conception rates in 2004. Clearly the points representing the providers do not move: the choice of approach only determines where the control limits lie. Ignoring the multiple testing issue for the moment, we use arbitrary one-sided p-value thresholds of 0.025 and 0.005 to determine the position of these limit lines. The estimate of ρ here is equal to 0.60, although, as noted above, this value is dependent on the particular risk-adjustment model used.

The first funnel plot highlights many providers as being potentially unusual, since the null model ignores the observed overdispersion around the mean. Use of the second approach leads to a substantially wider funnel. A few Local Authorities still lie beyond the limit lines in plot 2, suggesting that there

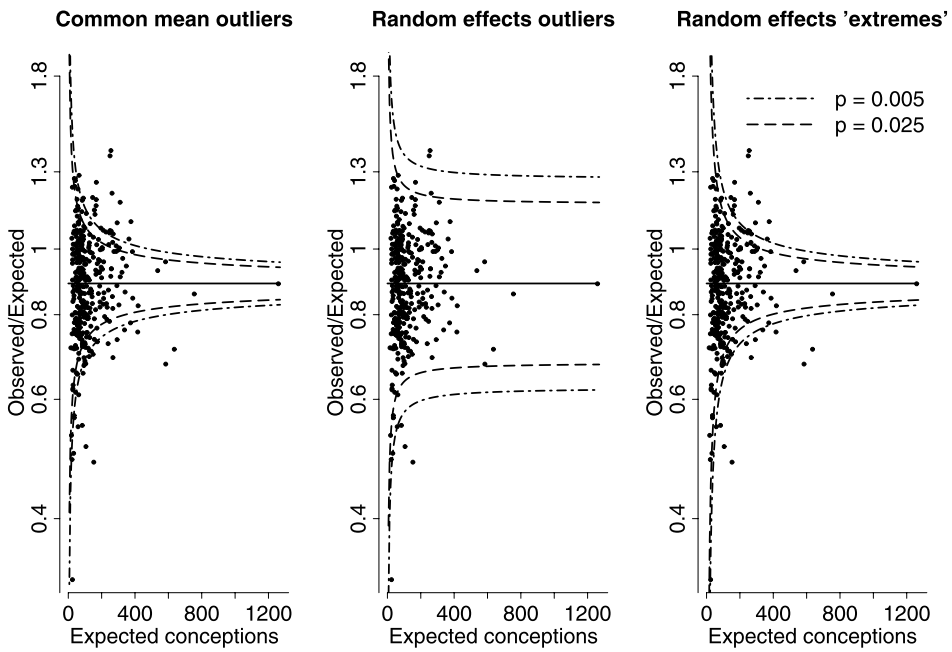

Figure 1. Teenage conception rates in Local Authorities, 2004: three possible approaches to identifying interesting providers. Results based on a normal approximation to the distribution of the log relative risks and empirical Bayes estimation ($\hat{\mu} = -0.13$, $\hat{\tau}^2 = 0.02$, $\bar{\sigma}^2 = 0.01$, $\hat{\rho} = 0.60$).

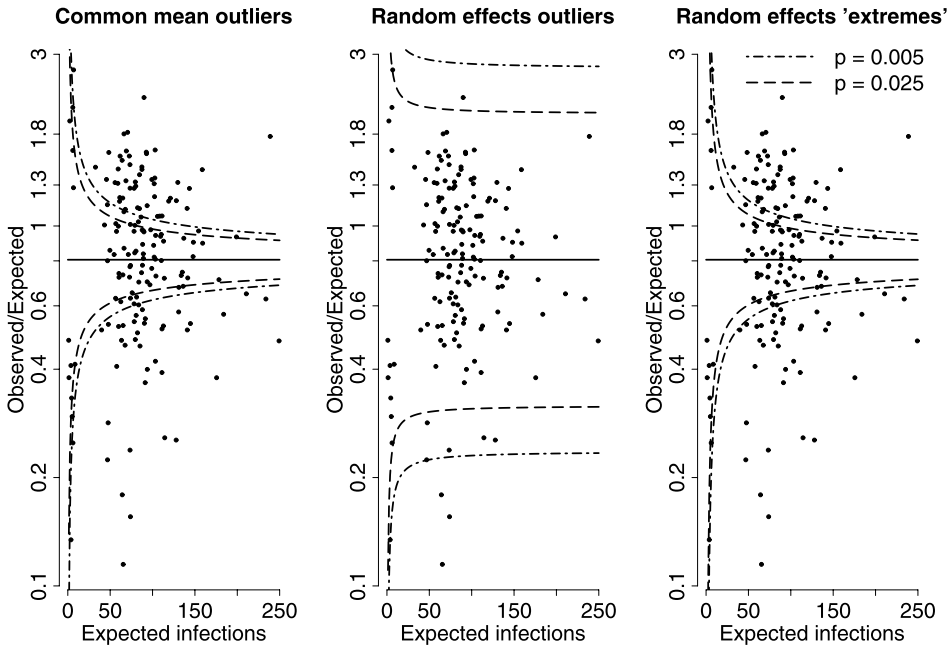

Figure 2. *Clostridium difficile* rates in NHS Trusts, October–December 2006: three possible approaches to identifying interesting providers. Results based on a normal approximation to the distribution of the log relative risks and empirical Bayes estimation ($\hat{\mu} = -0.22$, $\hat{\tau}^2 = 0.23$, $\bar{\sigma}^2 = 0.04$, $\hat{\rho} = 0.87$).

might be a few genuine outliers to the random effects distribution. Of course in practice the p-value threshold should be more carefully determined to account for the multiple testing problem. Note further that this analysis is slightly conservative since the parameters of the random effects distribution were not estimated using robust methods.

The third funnel is seen to be similar to the first, with only slightly fewer providers lying beyond the limit lines: with $\hat{\rho} = 0.60$, for some providers we can be fairly confident about which half of the distribution they lie in.

4.2 *Clostridium difficile* in NHS Trusts

Next consider rates of the infection *Clostridium difficile* in $m = 163$ NHS Trusts during the quarter October to December 2006. The between-Trust variability is large ($\hat{\tau}^2 = 0.23$), and ρ is estimated to be 0.87. This large ρ is not surprising given the infectious nature of *C. difficile* and the severely limited risk adjustment which we have employed here (Jones and Spiegelhalter 2009).

Figure 2 demonstrates the large number of outliers relative to the common mean model. In contrast, the second funnel is very wide, as the vast majority of Trusts are accommodated by the random effects model with large τ. The third funnel is almost identical to the first, providing practically no adjustment for the overdispersion, so that large providers will tend to signal using this approach whether they are interesting or not.

4.3 Mortality Following Heart Surgery in New York State Hospitals

Finally, we consider rates of mortality following CABG surgery in $m = 37$ New York State hospitals in 2003. Unlike the two previous examples, in which our own simple risk-adjustment has been employed using limited available information, the E_i's for these rates are provided by the New York State Department of Health (2009) and have been calculated using a sophisticated patient-level model. As a result, we would expect the observed relative risks to be relatively homogeneous. Indeed, since $\bar{\sigma}^2$ is 0.21, large relative to $\hat{\tau}^2 = 0.04$, ρ is estimated to be 0.17, much smaller than in the previous two examples.

Figure 3 shows that the second funnel for these data is not much wider than the first, as seems reasonable since there is little overdispersion. However, the third funnel is very wide: since the providers' rates are all similar to each other, we cannot be at all sure about which half of the distribution each one lies in. In particular, note that one hospital lies below the outer limit lines in the first and second funnels but is accommodated by the third.

We note that this is the same dataset as used by Racz and Sedransk (2010), although the particular data they used were from earlier years. It is the third funnel plot which is analogous to Racz and Sedransk's (2010) two random effects approaches. Our first two worked examples have demonstrated that, when substantial overdispersion to the common mean model exists, this method will tend to identify too many providers as unusual. However, Racz and Sedransk (2010) argued that their random effects methods instead identified too *few* "outliers" and as such that random effects models are inappropriate. Figure 3 clearly demonstrates that, for such a thoroughly risk-adjusted dataset, this is in fact the case: the authors would have identified more units as unusual if they had used approach 2, which correctly identifies outliers to the random effects distribution. We emphasize further that failing to detect truly unusual units should

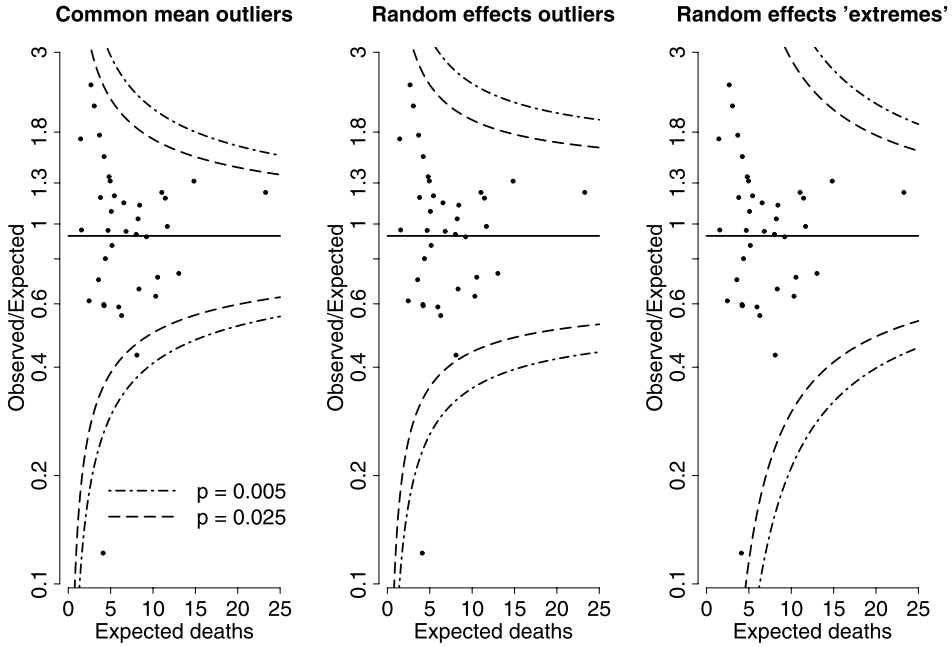

Figure 3. Mortality rates following CABG in New York State hospitals, 2003: three possible approaches to identifying interesting providers. Results based on a normal approximation to the distribution of the log relative risks and empirical Bayes estimation ($\hat{\mu} = -0.09, \hat{\tau}^2 = 0.04, \bar{\sigma}^2 = 0.21, \hat{\rho} = 0.17$).

not be the only concern when deciding upon a provider profiling scheme: false positives must be considered, since wrongly labeling a hospital as poor could have important consequences.

5. DISCUSSION

Hierarchical models have become widely used in performance monitoring, and have many attractive properties. The use of such models need not be computationally demanding: using an EB approach, as in this article, shrinkage estimates and associated error terms can be computed easily, even using a spreadsheet. For large m, the hyperparameters can be estimated quite precisely, so that the results will tend to be very similar to those from a fully Bayesian analysis.

Despite such models now being used frequently, there clearly remains some confusion about how to identify "unusual" performance based upon them. Using a two-level normal model, we have demonstrated an important distinction between identifying outliers and extremes. In practice, analysts must first decide between a hypothesis testing or estimation strategy, as discussed in Section 1, and be very clear upon this when summarizing results.

If following a hypothesis testing strategy, a simple random effects model such as the one-way ANOVA or Poisson–gamma model should suffice, and approach 2 should be used to identify outliers. As we have shown, this is in accordance with Goldstein's (2003) advice that "diagnostic" standard errors should be used when examining residuals to check for outliers. Care is necessary since statistical software may instead report "comparative" standard errors by default: if treating these as if they were diagnostic, analysts will inadvertently be identifying providers that are only "above or below average." Further, to remove the influence of any truly outlying providers on the null distribution, the hyperparameters should ideally be estimated using robust methods such as Winsorization (Spiegelhalter 2005b; Ohlssen, Sharples, and Spiegelhalter 2007). This method shrinks in more extreme values before incorporating them into the estimates of the hyperparameters.

If instead following an estimation strategy, then all providers should be accommodated by the random effects model. As we have discussed, it does not then make sense to check for outliers, other than to verify that the model fit is reasonable. More sophisticated models might be required to achieve this accommodation. For example, funnel 2 of Figure 1 indicated that the simple two-level normal model did not seem to accommodate all of the LAs. A heavier tailed distribution for the random effects, such as a t-distribution, or a mixture of normals might fit better (Ohlssen, Sharples, and Spiegelhalter 2007). Once a reasonable model has been fitted, approach 3 can be used to identify extreme providers, although comparisons should be made with an external target or an alternative quantile of the random effects distribution t using (10), not the population average μ. Funnel plot control limits can easily be positioned based on more appropriate posterior tail areas. Again, care should be taken with the wording, since providers lying beyond such lines are not necessarily statistical outliers.

For ease of exposition we have not discussed the related multiple testing problem in this article. Some authors have suggested that shrinkage estimation makes adjustments for multiple testing unnecessary (Thomas, Longford, and Rolph 1994; Morris and Christiansen 1996). However, from a hypothesis testing perspective, each of the approaches discussed requires classical p-values, and appropriate error rates such as the FDR should be controlled if using these p-values to identify "unusual" performance (Jones, Ohlssen, and Spiegelhalter 2008).

In general, if working in an estimation setting and simply *describing* rates in each provider, multiple testing might be ignored. But if making statements about the confidence with which particular providers are unusual, the properties of the identification process must be considered and the multiple testing accounted for, if only informally.

We have focused here primarily on analysis based on the one-way ANOVA model. However, given the general results (9) and (12), clearly the entire discussion is also valid for other models, including Poisson–gamma, if a normal approximation to the marginal and posterior distributions is used. Simpson et al. (2003) used such an approximation to place funnel plot limit lines, which were centered around the population average. They used an approximation to the posterior standard deviation (equivalently, the comparative standard error) to determine the exact locations of the limit lines, but we have argued in this article that this is not appropriate for identifying outliers.

For the Poisson–gamma model, the predictive and posterior distributions are in fact readily available and so the exact versions can be used without much difficulty. The distinction between general approaches to identifying unusual performance remains equally important: the marginal negative binomial distribution corresponding to this model can be used to identify outlying providers, while posterior tail areas based on the gamma posterior distribution of $r_i|O_i$ can be used to identify extremes.

Finally, we have emphasized throughout this article that all methods described are very general, rather than being tied exclusively to either a Bayesian or a classical framework.

[Received September 2010. Revised July 2011.]

REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300. [155]

Burgess, J. F., Christiansen, C. L., Michalak, S. E., and Morris, C. N. (2000), "Medical Profiling: Improving Standards and Risk Adjustment Using Hierarchical Models," *Journal of Health Economics*, 19, 291–309. [154,155]

Carlin, B. P., and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), New York: Chapman & Hall/CRC. [156]

Christiansen, C., and Morris, C. N. (1997a), "Hierarchical Poisson Regression Modeling," *Journal of the American Statistical Association*, 92, 618–632. [157]

——— (1997b), "Improving the Statisticial Approach to Health Care Provider Profiling," *Annals of Internal Medicine*, 127, 764–768. [155]

Clayton, D., and Kaldor, J. (1987), "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping," *Biometrics*, 43, 671–681. [157]

Coory, M., and Gibberd, R. (1998), "New Measures for Reporting the Magnitude of Small-Area Variation in Rates," *Statistics in Medicine*, 17, 2625–2634. [157]

Darlow, B. A., Hutchinson, J. L., Simpson, J. M., Henderson-Smart, D. J., Donoghue, D. A., and Evans, N. J. (2005), "Variation in Rates of Severe Retinopathy of Prematurity Among Neonatal Intensive Care Units in the Australian and New Zealand Neonatal Network," *British Journal of Ophthalmology*, 89, 1592–1596. [154,155]

Efron, B., and Morris, C. (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70, 311–319. [154]

Gelman, A., and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press. [156]

Goldstein, H. (2003), *Multilevel Statistical Models* (3rd ed.), London: Edward Arnold. [155,158,161]

Goldstein, H., and Spiegelhalter, D. J. (1996), "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance," *Journal of the Royal Statistical Society, Ser. A*, 159, 385–443. [154,157]

Greenland, S., and Robins, J. M. (1991), "Empirical-Bayes Adjustments for Multiple Comparisons are Sometimes Useful," *Epidemiology*, 2, 244–251. [154,156]

Hardy, R. J., and Thompson, S. G. (1998), "Detecting and Describing Heterogeneity in Meta-Analysis," *Statistics in Medicine*, 17, 841–856. [158]

Howley, P. P., and Gibberd, R. (2003), "Using Hierarchical Models to Analyse Clinical Indicators: A Comparison of the Gamma–Poisson and Beta–Binomial Models," *International Journal for Quality in Health Care*, 15, 319–329. [156]

James, W., and Stein, C. (1961), "Estimation With Quadratic Loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, pp. 361–379. [154]

Jones, H. E., and Spiegelhalter, D. J. (2009), "Accounting for Regression-to-the-Mean in Tests for Recent Changes in Institutional Performance: Analysis and Power," *Statistics in Medicine*, 28, 1645–1667. [154,159,160]

Jones, H. E., Ohlssen, D. I., and Spiegelhalter, D. J. (2008), "Use of the False Discovery Rate When Comparing Multiple Healthcare Providers," *Journal of Clinical Epidemiology*, 61, 232–240. [155,161]

Landrum, M. B., Normand, S.-L. T., and Rosenheck, R. A. (2003), "Selection of Related Multivariate Means: Monitoring Psychiatric Care in the Department of Veterans Affairs," *Journal of the American Statistical Association*, 98, 7–16. [155]

Lange, N., and Ryan, L. (1989), "Assessing Normality in Random Effects Models," *The Annals of Statistics*, 17, 624–642. [158]

Langford, I. H., and Lewis, T. (1998), "Outliers in Multilevel Data," *Journal of the Royal Statistical Society, Ser. A*, 161, 121–160. [158]

Lee, P. M. (1997), *Bayesian Statistics* (2nd ed.), London: Arnold. [158]

Louis, T. A. (1991), "Assessing, Accommodating and Interpreting the Influences of Heterogeneity," *Environmental Health Perspectives*, 90, 215–222. [154]

McPherson, K., Wennberg, J. E., Hovind, O. B., and Clifford, P. (1982), "Small-Area Variations in the Use of Common Surgical Procedures: An International Comparison of New England, England and Norway," *The New England Journal of Medicine*, 307, 1310–1314. [157]

Morris, C. N., and Christiansen, C. L. (1996), "Hierarchical Models for Ranking and for Identifying Extremes, With Applications," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, pp. 277–296. [154,155,161]

New York State Department of Health (2009), "Cardiovascular Disease Data and Statistics," available at *http://www.health.state.ny.us/statistics/diseases/cardiovascular/* (accessed August 2011). [160]

Normand, S.-L. T., Glickman, M. E., and Gatsonis, C. A. (1997), "Statistical Methods for Profiling Providers of Medical Care: Issues and Applications," *Journal of the American Statistical Association*, 92, 803–814. [154,155,157]

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007), "A Hierarchical Modelling Framework for Identifying Unusual Performance in Health Care Providers," *Journal of the Royal Statistical Society, Ser. A*, 170, 865–890. [155,156,161]

Racz, M. J., and Sedransk, J. (2010), "Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes," *Journal of the American Statistical Association*, 105, 48–58. [154,155,160]

Rubin, D. B. (1980), "Using Empirical Bayes Techniques in Law School Validity Studies," *Journal of the American Statistical Association*, 75, 801–816. [154]

Schulman, J., Spiegelhalter, D. J., and Parry, G. (2008), "How to Interpret Your Dot: Decoding the Message of Clinical Performance Indicators," *Journal of Perinatology*, 28, 588–596. [155,159]

Simpson, J. M., Evans, N., Gibberd, R. W., Heuchan, A. M., and Henderson-Smart, D. J. (2003), "Analysing Differences in Clinical Outcomes Between Hospitals," *Quality and Safety in Health Care*, 12, 257–262. [154,155,157,162]

Smits, J. M. A., Meester, J. D., Deng, M. C., Scheld, H. H., Hummel, M., Schoendube, F., Haverich, A., Vanhaecke, J., and van Houwelingen, H. C. (2003), "Mortality Rates After Heart Transplantation: How to Compare Center-Specific Outcome Data?" *Transplantation*, 75, 90–96. [154,155]

Spiegelhalter, D. J. (2005a), "Funnel Plots for Comparing Institutional Performance," *Statistics in Medicine*, 24, 1185–1202. [155,159]

——— (2005b), "Handling Over-Dispersion of Performance Indicators," *Quality and Safety in Health Care*, 14, 347–351. [154,156,161]

Teixeira-Pinto, A., and Normand, S.-L. T. (2008), "Statistical Methodology for Classifying Units on the Basis of Multiple-Related Measures," *Statistics in Medicine*, 27, 1329–1350. [155]

Thomas, N., Longford, N. T., and Rolph, J. E. (1994), "Empirical Bayes Methods for Estimating Hospital-Specific Mortality Rates," *Statistics in Medicine*, 13, 889–903. [154,161]

Tomberlin, T. J. (1988), "Predicting Accident Frequencies for Drivers Classified by Two Factors," *Journal of the American Statistical Association*, 83, 309–321. [154]

U.K. Care Quality Commission (2009), "Statistical Banding for Existing Commitment and National Priority Indicators, Annual Health Check 2008/2009," available at *http://www.cqc.org.uk/_db/_documents/Stats_banding_target_methods_20081126.pdf* (accessed August 2011). [155]