

Decision theory for comparing institutions

Nicholas T. Longford,
Imperial College, London, United Kingdom

Abstract

Various forms of performance assessment are applied to public service institutions, such as hospitals, schools, police units and local authorities. Difficulties arise in the interpretation of the results presented in some established formats because they require a good understanding and appreciation of the uncertainties involved. Usually the results have to be adapted to the perspectives of the users — managers of the assessed units, a consumer, or a central authority (a watchdog) that dispenses awards and sanctions. We present a decision-theoretical approach to these and related problems in which the perspectives are integrated in the analysis and its results are choices from a finite list of options (alternative courses of action).

Keywords: *analyst and client; outlier; performance assessment; separator; verdict; winner.*

N. T. Longford, Department of Medicine, Imperial College, London, UK.
Email: `n.longford@imperial.ac.uk`.

1 Introduction

Performance assessment of institutions (units) entails various comparisons of units on the basis of their outcomes recorded on a common scale and over a common period of time, such as a year. The forms of comparison include identifying the best and the worst units, classifying them into a small number of ordinal categories, comparing two units (which of them is superior), and ranking (compiling a league table) as their extreme form. The results of a performance assessment have nontrivial consequences on the units, especially those that are highlighted for unusual performance. They often spawn a variety of secondary analyses. The units' managers face a difficult task of drawing appropriate conclusions from the assessment and selecting from a set of contemplated courses of action or devising actions that best respond to the results [and to the uncertainties they involve](#). Goals related to future assessments, such as to move up the league table, to be classified more favourably, and the like, are often based on the assumptions of stationarity (that the other units will not perform differently) and they ignore the uncertainty associated with the statements made or implied by a performance assessment report.

The general subject of performance assessment has received a lot of high-profile attention recently ([1] and references therein). The solutions proposed and implemented are based on hypothesis testing [and confidence intervals](#), and their adaptations for binary choices, such as z-scores and funnel plots, [2]. [They are an outgrowth of methods for comparing two units](#) [3]. In this paper, we present decision-theoretical solutions to several common problems in performance assessment or, more generally, in the analysis of a collection of units $k = 1, \dots, K$, each with a set of elementary outcomes y_{ik} , $i = 1, \dots, n_k$, which are summarised by respective statistics $\hat{\theta}_k$. We assume that each $\hat{\theta}_k$ is an unbiased estimator of a characteristic of the performance, and the purpose is to compare their targets θ_k .

The key innovation of our approach is integration of the perspective (value judgements and preferences) and of the consequences of incorrect statements made in the analysis. In the ideal implementation of our approach, the result of the analysis is the course of action that is optimal in a well defined sense. It requires no interpretation. The solution we propose is not without difficulties — it requires input additional to the

data used by established methods. Obtaining it, by elicitation, is often a contentious exercise, but some leeway is afforded by specifying so called plausible ranges instead of values of the elicited parameters. A similar approach is applied to educational evaluation, [4]. It can be motivated as a sensitivity analysis, evaluating the results for a range of plausible settings. For general background to decision theory, see [5]–[7].

The next five sections deal with distinct tasks related to performance assessment: comparing two units’ averages or a unit’s and the overall (national) averages, classifying the units to a small number of a priori defined ordinal categories, identifying exceptional units, the best and worst units, and units that are almost as good as the best or almost as poor as the worst. They are adapted to the context of health-care institutions from [4], where an application to educational research is presented. These methods involve no substantially new developments, but novel is their application to performance assessment in health care in Section 7, where we analyse data from the National Neonatal Audit Programme (UK). We compare the rates of compliance with an established standard in the neonatal care networks (groups of hospitals) in England, Scotland and Wales in 2015. Section 8 deals with the corresponding problem for hospitals. It presents an application involving beta distributions. The concluding section summarises the strengths of the presented approach, discusses its full potential and addresses some peripheral issues. Difficulties with implementing our general proposal are also discussed.

2 Comparing two units

We outline the proposed approach on the elemental problem of deciding which unit has a better (or worse) outcome. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be estimators of θ_1 and θ_2 , with respective sampling distributions $\mathcal{N}(\theta_1, \sigma_1^2)$ and $\mathcal{N}(\theta_2, \sigma_2^2)$. We have to choose one of three contemplated statements:

1. $\theta_1 \prec \theta_2$;
2. $\theta_1 \simeq \theta_2$;
3. $\theta_1 \succ \theta_2$.

Statement 2 stands for ‘ θ_1 and θ_2 differ insubstantially, or not at all.’ The inequalities $\prec$ and $\succ$ in respective statements 1 and 3 are defined so that the three statements are

mutually exclusive (disjoint) and exhaustive (they cover all the possibilities). Details of these definitions are discussed below. We distinguish between the *state* — one of the statements 1, 2 or 3 which corresponds to the unknown reality, and a *verdict* — a data-based choice of one of these statements.

Let $\Delta\theta = \theta_1 - \theta_2$. In an established approach, the hypothesis $H_0 : \Delta\theta$ is tested and verdict 1 or 3 is issued if the hypothesis is rejected; the choice between the two verdicts is based on the sign of $\Delta\hat{\theta}_1$. [The method is commonly abused by issuing](#) verdict 2 if the hypothesis is not rejected because the verdict is not supported by any evidence. Also, the qualifier ‘insubstantial’ (or approximate) in statement 2 is rarely specified with any rigour. For a cogent discussion of deficient practice of hypothesis testing, see [8]. [For our perspective, see \[9\] and \[10\].](#)

In the solution we propose, we partition the real interval $(-\infty, +\infty)$ to three subsets that correspond to the three states. For each state, a different course of action would be applied. The three or another finite number of verdicts may be associated with any partitioning of $(-\infty, +\infty)$, but here we consider only partitioning to three intervals defined by two thresholds $T_1 \leq T_2$. This includes settings with only two intervals, for which $T_1 = T_2$. The partitioning is usually symmetric, $T_1 = -T_2$, but our development is not restricted to this condition.

The consequences of issuing an incorrect verdict, that is, a verdict v different from the state V , are quantified by means of losses. We refer to the unit of loss as a lossile, and treat it like a unit of currency (for error). An essential property of the loss is linearity: the loss of l_1 lossiles in one instance and of l_2 lossiles in another is for all purposes equivalent to the loss of $l_1 + l_2$ lossiles in a single instance. When verdict v is issued but the state is V , the loss of L_{vV} lossiles is incurred. Correct verdict entails no loss; $L_{VV} = 0$ for all V .

Loss is a random variable, so it is natural to consider its expectation. [In the Bayesian paradigm, the expectation is evaluated with respect to the posterior distribution of the focal parameter \(\$\Delta\theta\$ \). Below we refer to posterior probabilities, distributions and expectations, but we emphasize that the methods are equally applicable in a frequentist analysis by adopting the fiducial argument, \[11\]–\[13\]. See Longford \[10\] for a recent application. Lindley \[14\] established a correspondence between posterior and fiducial distributions.](#)

Prior to realising the data, $\Delta\theta$ is a fixed (unknown) quantity and its estimator $\Delta\hat{\theta}$ is a random variable. Once $\Delta\hat{\theta}$ is realised it becomes an estimate, with a fixed value, and $\Delta\theta$ becomes a random variable. When $\Delta\hat{\theta} = \Delta\theta + \varepsilon$ is originally a normally distributed random variable, [Bayes theorem with a flat prior can be interpreted as](#) $\Delta\theta = \Delta\hat{\theta} - \varepsilon$, and hence $\Delta\theta \sim \mathcal{N}(\Delta\hat{\theta}, \sigma_1^2 + \sigma_2^2)$, assuming that prior to their realisation $\hat{\theta}_1$ and $\hat{\theta}_2$ were independent. We consider the expected loss with respect to this [posterior](#) distribution. Denote $\Delta\sigma^2 = \sigma_1^2 + \sigma_2^2$ and $\Delta\sigma = \sqrt{\Delta\sigma^2}$.

If we choose verdict 1 ($\Delta\theta < T_1$), then the expected loss is

$$\begin{aligned} Q_1 &= L_{12} \int_{T_1}^{T_2} \phi(x; \Delta\hat{\theta}, \Delta\sigma^2) dx + L_{13} \int_{T_2}^{+\infty} \phi(x; \Delta\hat{\theta}, \Delta\sigma^2) dx \\ &= L_{13} - L_{12} \Phi(T_1; \Delta\hat{\theta}, \Delta\sigma^2) - (L_{13} - L_{12}) \Phi(T_2; \Delta\hat{\theta}, \Delta\sigma^2), \end{aligned}$$

where $\phi(x; \mu, \sigma^2)$ is the density and $\Phi(x; \mu, \sigma^2)$ the normal distribution function of $\mathcal{N}(\mu, \sigma^2)$. We drop the arguments μ and σ^2 when $\mu = 0$ and $\sigma^2 = 1$. Thus, $\phi(x; \mu, \sigma^2) = \phi\{(x - \mu)/\sigma\}/\sigma$ and $\Phi(x; \mu, \sigma^2) = \Phi\{(x - \mu)/\sigma\}$. Denote $u_h = (T_h - \Delta\hat{\theta})/\Delta\sigma$, $h = 1, 2$, so that

$$Q_1 = L_{13} - L_{12} \Phi(u_1) - (L_{13} - L_{12}) \Phi(u_2).$$

Verdicts 2 ($T_1 < \Delta\theta < T_2$) and 3 ($\Delta\theta > T_2$) are associated with respective expected losses

$$\begin{aligned} Q_2 &= L_{23} + L_{21} \Phi(u_1) - L_{23} \Phi(u_2) \\ Q_3 &= (L_{31} - L_{32}) \Phi(u_1) + L_{32} \Phi(u_2). \end{aligned}$$

We choose the verdict v for which Q_v is smaller than for the two alternatives. In any reasonable setting, larger error is associated with greater loss: $L_{13} > L_{23}$, $L_{13} > L_{12}$, $L_{31} > L_{32}$ and $L_{31} > L_{21}$. We assume these inequalities, unless stated otherwise. However, L_{13} and L_{31} , as well as similar pairs, may differ. The assumption of additivity, that $L_{13} = L_{12} + L_{23}$ and $L_{31} = L_{32} + L_{21}$ is often, but not always, reasonable.

The properties of Q_1 , Q_2 and Q_3 as functions of $\Delta\hat{\theta}$ are easy to explore. In particular, $Q_1(\Delta\hat{\theta})$ is increasing, with limits of zero as $\Delta\hat{\theta} \rightarrow -\infty$ and L_{13} as $\Delta\hat{\theta} \rightarrow +\infty$, and $Q_3(\Delta\hat{\theta})$ is decreasing from L_{31} at $-\infty$ to zero at $+\infty$. Therefore verdict 1 is inappropriate for sufficiently large $\Delta\hat{\theta}$ and verdict 3 for sufficiently small $\Delta\hat{\theta}$. The balance

equation for verdicts 1 and 3, $Q_1 = Q_3$, has a unique solution that can be found by numerical methods. The solution is zero in some symmetric settings.

The behaviour of $Q_2(\Delta\hat{\theta})$ depends on the relative sizes of L_{12} and L_{23} and the distance $T_2 - T_1$, since

$$\frac{\partial Q_2}{\partial \Delta\hat{\theta}} = \frac{1}{\Delta\sigma} \{L_{23}\phi(u_2) - L_{21}\phi(u_1)\}$$

may have either sign at the thresholds T_1 and T_2 .

Figure 1 presents an example of the three expected loss functions for the setting with $T_1 = -1$, $T_2 = 1$, $\Delta\sigma = 0.39$ and losses $L_{12} = 1.0$, $L_{13} = 2.5$, $L_{21} = L_{23} = 0.5$, $L_{31} = 2.5$ and $L_{32} = 1.0$, described by the loss matrix

$$\begin{pmatrix} 0.0 & 1.0 & 2.5 \\ 0.5 & 0.0 & 0.5 \\ 2.5 & 1.0 & 0.0 \end{pmatrix}. \quad (1)$$

The diagonal of this matrix corresponds to correct verdicts which are associated with no loss.

FIGURE 1

The borderline between verdicts 1 and 2 is at $B_{12} = -1.165$ and between verdicts 2 and 3 at $B_{23} = 1.165$. Verdict 1 is issued when $\Delta\hat{\theta} < B_{12}$, verdict 2 when $B_{12} < \Delta\hat{\theta} \leq B_{23}$ (grey shading in the diagram) and verdict 3 when $\Delta\hat{\theta} > B_{23}$. The expected loss with this selection of the verdict is drawn by full solid line. This function attains its twin maxima at B_{12} and B_{23} , equal to 0.336. It converges to zero as $\Delta\hat{\theta} \rightarrow \pm\infty$. The local minimum of Q_2 attained at $\Delta\hat{\theta} = 0$ is positive, albeit very small, equal to 0.005. The domain of verdict 2, the interval (B_{12}, B_{23}) in Figure 1, is wider than the inter-threshold range (T_1, T_2) . This is a consequence of the balance of the losses $L_{12} = 1.0$ and $L_{21} = 0.5$, which can be interpreted as an aversion to the (inequality) verdicts 1 and 3. For sufficiently large L_{21} and L_{23} , $Q_2(0) > Q_1(0)$ and $Q_2(0) > Q_3(0)$; then verdict 2 is not issued for any value of $\Delta\hat{\theta}$. For example, verdict 2 is not issued for any $\Delta\hat{\theta}$ when $L_{21} = L_{23} = 2.0$, $L_{12} = L_{32} = 1.0$, $L_{13} = L_{31}$ and $\Delta\sigma = 2$. In this case, $Q_2(0) = 1.234$ whereas $Q_1(0) = Q_3(0) = 1.154$. In general, verdict 2 has a domain for sufficiently small $\Delta\sigma^2$, when $\Phi(u_2) - \Phi(u_1)$ is sufficiently large.

As the inter-threshold range is made narrower (and other parameters are held fixed), the domain of verdict 2 also becomes narrower. For $T_1 = T_2 = 0$, which has some affinity

to the setting of a null hypothesis, verdict 2 is ruled out. We motivate this as follows. We would choose the alternative $\theta \neq 0$, without any data analysis, because the solitary zero is a very poor bet against the continuum of values of θ , uncountably many of them in arbitrarily close distance from zero. The only exception might be a setting in which there is prior information in favour of the exact zero. But in this context even 10^{-20} differs from zero and belongs to the alternative.

2.1 Multiplicity

Unlike in hypothesis testing, making many comparisons by our approach raises no issues of having to adjust for multiplicity, [15] and [16], because the complex calculus of probabilities (for events that are not independent) is replaced by a simple linear calculus for (expected) loss — the losses, or the expected losses, are simply added up. This advantage does not come for free; the losses have to be defined on a scale that is linear, so that they could be treated as a currency for error. As more comparisons are made, the compound expected loss is bound to be greater. It may be useful to define a ‘fee’ for a comparison, to control the incentive to conduct comparisons without any restraint. This fee could be specific to the comparison made, reflecting its a priori importance. The units for the fee have to be lossiles, so that the investment made in a comparison can be related to the (expected) loss.

3 Classification

The task dealt with in the previous section is a special case of assigning the units to a set of ordinal categories. We partition the real axis according to a set of cut-points $C_1, \dots, C_{J-1}$, and seek an assignment of each unit k to one of the states (classes) $1, \dots, J$; state j is for units k for which $C_{j-1} \leq \theta_k < C_j$, where $C_0 = -\infty$ and $C_J = +\infty$. For each unit k we evaluate the posterior probabilities

$$r_{k,j} = \Phi\left(\frac{C_j - \hat{\theta}_k}{\sigma_k}\right) - \Phi\left(\frac{C_{j-1} - \hat{\theta}_k}{\sigma_k}\right),$$

$j = 1, \dots, J$. The losses are defined by a $J \times J$ matrix $\{L_{vV}\}$ for verdicts v and states V . The simplest example of such a matrix is given by $L_{vV} = |v - V|$, $1 \leq v, V \leq J$, counting the number of cut-points that separate verdict v from state V . More generality

is afforded by introducing a loss ratio $R > 0$ and specifying losses that satisfy the identity $L_{vV} = RL_{Vv}$ for $v > V$. Further, some cut-points may be regarded as more (or less) important and more (or less) than one lossile (R lossiles) is charged for an error involving them. An example of a loss matrix with these features is

$$\mathbf{L} = \begin{pmatrix} 0 & 4 & 8 & 20 \\ 1 & 0 & 4 & 16 \\ 2 & 1 & 0 & 12 \\ 5 & 4 & 3 & 0 \end{pmatrix}.$$

Here $J = 4$, $R = 4$ and the cut-point between classes 3 and 4 is regarded as three times more important than between classes 1 and 2 or 2 and 3. There is a strong aversion to understatement; for example $L_{14} = 20$ is $R = 4$ times greater than $L_{41} = 5$. Thus, erring on the side of overstatement is a good strategy, but this has to be done in good measure because issuing verdict 4 without any discrimination would be associated with excessive expected loss.

4 Exceptional units

Units whose performance is characterised by their values θ_k as very good or very poor are often identified by means of criteria for (statistical) outliers, [17] and [18]. A unit is regarded as an outlier if by removing it the remainder of units would fit the adopted model much better. A similar definition is used for a set of units. A unit may on its own be not an outlier, even when it is an element of a set of outliers. We want to refrain from assuming a model, especially when it plays a key role in as contentious an analysis as identifying units for a special treatment, such as an award or a sanction. Further ambiguity arises when the values of θ_k and $\hat{\theta}_k$ are confused — outliers found among $\hat{\theta}_k$ are put forward as outliers among θ_k .

We address these concerns by the following approach. We specify a value $G > 0$, called the *gap*, and declare a set of units $\mathcal{X}$ as exceptionally poor if

$$\min_{k \notin \mathcal{X}} \theta_k - \max_{k \in \mathcal{X}} \theta_k \geq G; \tag{2}$$

that is, when there is a gap at least G wide between the exceptional and the other (mainstream) units. The equation for exceptionally high values is obtained by exchanging the roles of $\mathcal{X}$ and its complement, or by applying the definition in (2) to

$-\theta_k$. The location (centre) Λ of the interval $(\Lambda - \frac{1}{2}G, \Lambda + \frac{1}{2}G)$ is not specified, only its width is given. The size of the set $\mathcal{X}$ is not specified either, but it is understood that it is less than half the sample size.

We evaluate the posterior probability of the event given by (2) for a range of locations of the gap and assess informally whether the set $\mathcal{X}$ is non-empty. Note that this process may establish whether there are some exceptional units; it does not identify the actual units. *Some ambiguity may arise for a unit with large σ_k , when the value of the estimate $\hat{\theta}_k$ does not provide a clear indication as to whether θ_k lies on one side of the gap or the other.* The posterior probability of a gap G centred at Λ is

$$P(\Lambda, G) = \prod_{k=1}^K \left\{ 1 - \Phi\left(\Lambda + \frac{1}{2}G; \hat{\theta}_k, \sigma_k\right) + \Phi\left(\Lambda - \frac{1}{2}G; \hat{\theta}_k, \sigma_k\right) \right\}. \quad (3)$$

Evaluating this probability for several values of Λ raises no concerns related to capitalisation on chance (multiplicity) because we want to declare a set of units as exceptional if any location Λ can be found for which $P(\Lambda, G)$ is substantial. A threshold P^* for P can be specified directly. If $\max_{\Lambda} P(\Lambda, G) > P^*$, then units k with $\hat{\theta}_k < \Lambda - \frac{1}{2}G$ are declared as exceptionally poor if the number of such units is smaller than $\frac{1}{2}K$; otherwise units with $\hat{\theta}_k > \Lambda + \frac{1}{2}G$ are declared as exceptionally good. We refer to the interval $(\Lambda^* - \frac{1}{2}G; \Lambda^* + \frac{1}{2}G)$ or to Λ^* for which $P(\Lambda, G)$ attains a local maximum that is substantial as a *separator*.

The term ‘substantial’ is qualified by decision-theoretical considerations. Suppose inappropriate declaration of a separator results in the loss of one lossile, and failure to declare a separator is associated with the loss of R lossiles. Then the respective expected losses are $1 - P$ and RP , where $P = P(\Lambda, G)$ is evaluated at a local maximum. By comparing the two expected losses we obtain the criterion $P > 1/(R + 1)$, so we set $P^* = 1/(R + 1)$.

The gap G and the loss ratio R are set by elicitation from experts. Ambiguity over their values can be addressed by specifying plausible ranges of their values, (G_L, G_H) and (R_L, R_H) , and solving the problem on a fine grid of points in the rectangle they define. This can be usually reduced to solving it for the four vertices of the rectangle. If the four solutions coincide, then the common solution defines an *unequivocal* verdict on every unit. Otherwise ambiguity (*impasse*) arises about one or several units because for some plausible configurations of G and R they would be declared as exceptional

and for others they would not be. This highlights the role of elicitation. On the one hand, plausible ranges should be as narrow as possible. On the other, their integrity is maintained only if all values of the parameters outside the plausible ranges can be ruled out.

5 Winners and losers

The posterior probability that unit k has the highest (best) value θ_k among the K units is obtained by integrating the conditional probability of this event given θ_k :

$$P(\theta_k \geq \max_{k'} \theta_{k'}) = \int_{-\infty}^{+\infty} \frac{\phi(z_k)}{\Phi(z_k)} \prod_{k'=1}^K \Phi(z_{k'}) dx, \quad (4)$$

where $z_k = (x - \hat{\theta}_k)/\sigma_k$ is a function of x . This probability, denoted by P_k , is evaluated numerically. We treat the issue of declaring a unit as a winner as a decision-theoretical problem. Suppose inappropriate declaration of a unit is associated with the loss of one lossile, whereas a failure to declare a (worthy) unit as the winner results in the loss of R lossiles. Then the expected loss related to inappropriate declaration of unit k is $1 - P_k$ and for the failure to declare unit k when it is the winner it is RP_k . Therefore, we declare unit k as a winner if $P_k > 1/(R + 1)$, and leave it out of the list of winners otherwise. Define the balance for unit k as $B_k = (R + 1)P_k - 1$, so that unit k is declared as a winner if $B_k > 0$. For sufficiently small R , k is not declared a winner. If $P_k > 0$, then k is a winner for sufficiently large R .

Although we make assumptions (of normality and known variance) about the sampling distribution of each estimator $\hat{\theta}_k$, we make no assumptions about the distribution of the values θ_k , which are required in a hierarchical model. In particular, we do not assume that this distribution is (absolutely) continuous, and so we do not rule out a tie, in which case there may be multiple winners. However, we draw a distinction between being a winner (a state V) and being declared a winner (a verdict v); v is error-prone.

With the assumption of continuity of the distribution underlying θ_k , ties are ruled out, so there is only one winner. No winner or several winners could be declared nevertheless. In these cases, the loss is positive with probability 1. However, the expected loss may be smaller than for declaring a single winner, which may incur zero loss with positive probability.

The preference for declaring a single winner can be implemented by introducing a sequence of losses (penalties) S_m , $m = 0, 1, \dots$, that apply when $m = |\mathcal{M}|$ winners are declared. Thus $B_{\mathcal{M}} = \sum_{k \in \mathcal{M}} B_k + S_{|\mathcal{M}|}$ is the corresponding balance. Only non-decreasing sequences $S_1, S_2, \dots$ are meaningful, with $S_1 = 0$, although $S_0 > 0$ can be set as the loss when no winner is declared. Suppose a set of winners $\mathcal{M}$ is obtained by minimising the expected loss with no penalty for sample size; $S_0 = S_1 = \dots = 0$. If $B_{\mathcal{M}} > 0$, we keep discarding from $\mathcal{M}$ the unit with the largest value B_k until we obtain a set $\mathcal{M}'$ with $B_{\mathcal{M}'} < 0$. Declaring a single winner can be enforced by setting S_0 and S_2 to sufficiently large values.

We may impose on the winner k a condition additional to being superior to the other units, namely that $\theta_k > U$, where U is an externally defined standard. This is implemented by defining a pseudo-unit $K + 1$ with $\theta_{K+1} = U$ and $\sigma_{K+1}^2 = 0$. If in the contest of these $K + 1$ units no winner is declared, or unit $K + 1$ is the only declared winner, then, in effect, no winner is declared. If a penalty sequence S_m is defined, it has to be adjusted to reflect the special status of pseudo-unit $K + 1$. There should be no penalty if no winner, not even the pseudo-unit, is declared, nor if the pseudo-unit is declared a winner together with a sole unit $k \leq K$.

6 Being one of the best

Unit k is said to be *one of the best*, if its value of θ_k is sufficiently close to $\theta_{\max} = \max_{k'} \theta_{k'}$. We adapt the method introduced in Section 4 by defining a gap G and regarding unit k as one of the best if $\theta_k > \theta_{\max} - G$. For a given location (centre) Λ , we evaluate the probability that $\theta_{k'} < \Lambda + \frac{1}{2}G$ for every unit k' and that $\theta_k > \Lambda - \frac{1}{2}G$ for the focal unit k . The probability of this event is

$$W_k(\Lambda, G) = \left\{ 1 - \frac{\Phi\left(\Lambda - \frac{1}{2}G; \theta_k, \sigma_k\right)}{\Phi\left(\Lambda + \frac{1}{2}G; \theta_k, \sigma_k\right)} \right\} \prod_{k'=1}^K \Phi\left(\Lambda + \frac{1}{2}G; \theta_{k'}, \sigma_{k'}\right). \quad (5)$$

We choose the location Λ for which this probability attains its maximum and issue a verdict by relating it to an adopted loss ratio R . Just as in Section 4, no issues of multiplicity related to maximising $W_k(\Lambda, G)$ as a function of Λ arise here either. The optimal value of Λ can be found by a numerical method, such as the Newton-Raphson algorithm, although evaluation of W_k on a dense grid of values of Λ is a practical

alternative. It is easy to show that W_k is an increasing function of G ; by G we control the inclusiveness of the list of units declared as being among the best. Instead of specifying G upfront, we may search for the smallest value G_k for which unit k would be declared as being one the best. If this value G_k is plausible (small enough), then the declaration of unit k as being one of the best is warranted.

By dropping all factors in (5) that involve units $k' \neq k$ and exploiting the symmetry of the normal distribution, we obtain the inequality $W_k < 2\Phi(\frac{1}{2}G/\sigma_k) - 1$. It implies that the value of G has to be informed by the prevailing values of the sampling variation of $\hat{\theta}_k$. The method based on W_k has a simple adaptation for identifying units that are not much better than the worst unit. Verdicts on being above the national rate or another threshold (standard) can be derived similarly.

7 Application

The National Neonatal Audit Programme (NNAP) conducts an annual audit of the care for infants born preterm (with gestational age lower than 37 weeks) and with low birthweight (less than 1500 grams). In 2015, the care was provided in 177 participating neonatal care units (hospitals) in England and Wales, and part of Scotland. The units were organised in 16 geographically contiguous neonatal networks. The audit reports the estimates of the rates (percentages) of compliance with several well established practices and standards in neonatal care. These estimates, evaluated for units and networks, are based on the data extracted from the National Neonatal Research Database (NNRD).

The first of the eleven standards (measures) that are subject to the audit is concerned with timely first measurement of body temperature of the newborn and keeping the infant warm immediately after birth. According to the standard established by NNAP, every infant born at a gestational age lower than 32 weeks should have its temperature taken within 60 minutes of birth, and this temperature should be within the ‘healthy’ range $36.5 - 37.5^\circ\text{C}$. The treatment of an infant is in compliance if both of these conditions (timeliness and range) are satisfied. Compliance, or its failure, is defined for every infant in the specified population. Some records are incomplete. Compliance is classified as missing if the data entry is missing for both conditions or

if one of the two conditions is recorded as having been satisfied and the record for the other condition is missing. Compliance is classified as a failure if at least one of the two conditions is recorded as not satisfied.

We focus on the network-level rates because care for the mother, especially in the last few days before giving birth, may be provided in one hospital, and delivery takes place in another. In some cases, the newborn is transferred from the unit of birth to another unit for specialised care. Most of these transfers take place within a network. We keep the networks (and hospitals in Section 8) anonymous; their identity is not relevant to the analysis.

In 2015, 177 neonatal care units from 16 networks contributed data to NNRD. Their total caseload in the studied population was 7783 infants. Compliance is not established for 26 infants (0.3%). The highest count of such cases in a network is 5, and it accounts for 0.8% of its infants. To simplify the discourse, we ignore the issue of missing data. A simple way of dealing with it is by applying the two kinds of extreme imputation — to set all missing entries to Y (complying with the standard) in one instance and to set them to N (failing to comply) in the other — and reporting the corresponding sets of results. This is sufficient for estimating the rates of compliance with the standard for temperature, but not for some other standards. A principled approach to this problem is by multiple imputation, [19] and [20].

7.1 Network vs. country

Figure 2 displays the estimated network-level rates of compliance ($\hat{\theta}_k$, $k = 1, \dots, K = 16$) in a caterpillar plot. The caseloads of the networks are printed at the top margin, and the national (overall) caseload, with the 26 missing entries removed, at the horizontal dashes that mark the estimated national rate, 58.4%. The standard error of this rate is estimated by 0.56%. The 95% confidence interval for the national rate is indicated by the darker grey strip around the estimated rate. The diagram suggests that the network-level rates vary substantially, and that there may be one network with exceptionally low and one with exceptionally high rate.

FIGURE 2

We compare each network's rate with the overall (national) rate. The variance of the contrast of the estimators of a network's and the national rates is

$$\Delta\sigma_k^2 = \frac{\theta_k(1-\theta_k)}{n_k} - \frac{2\theta_k(1-\theta_k)}{n} + \sum_{k'=1}^K \frac{n_{k'}}{n^2} \theta_{k'}(1-\theta_{k'}) , \quad (6)$$

where n_k is the number of infants, $n = n_1 + \dots + n_K$; $\hat{\theta}_k = m_k/n_k$, where m_k is the number of cases of compliance in network k and θ_k is the underlying rate. In the expression for $\Delta\sigma_k^2$, we replace each instance of θ_k and $\theta_{k'}$ by its estimate. This introduces only slight additional uncertainty because $\theta(1-\theta)$ is a flat function of θ in the range of the observed rates (40–75%). The second and third terms in (6) can be ignored whenever $n_k \ll n$ and θ_k is not very distant from 0.5.

We assume that a difference within $\pm 2.5\%$ is insubstantial, $T_1 = -0.025$ and $T_2 = 0.025$; this is indicated in Figure 2 by the faint-grey horizontal strip. We set the loss matrix to

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 5 \\ 2 & 0 & 2 \\ 5 & 1 & 0 \end{pmatrix} . \quad (7)$$

That is, failure to detect a substantial difference incurs twice as high loss as detecting it incorrectly, and confusing the two inequalities is regarded as a grave error, incurring the loss of 5 lossiles.

The expected losses for the three options are listed in Table 1, with the smallest value for each network underlined. Networks 1–7, with the lowest rates, are issued verdict 1 ($\theta_k \prec \theta$), networks 8–10 verdict 2 ($\theta_k \simeq \theta$), and networks 11–16, with the highest rates, are assigned verdict 3 ($\theta_k \succ \theta$).

TABLE 1

There may be some contention about the verdicts for networks 7 and 10, because the second smallest expected losses for them (for respective verdicts 2 and 3) are only slightly greater. Relatively small changes in matrix $\mathbf{L}$ may result in a different verdict for one or the other network. We say that the verdicts for these two networks are *equivocal*. The uncertainty and ambiguity about the 16 verdicts is summarised by the total expected loss, equal to 4.73 lossiles. By definition, such a summary is meaningful only if the loss used is a good (linear) currency for error. The expected loss is distributed

unevenly across the districts. It is smaller than 0.1 for networks 1, 2, 3, 15 and 16, whereas units 7 and 11 account for over 30% of the total.

Ambiguity about the loss matrix $\mathbf{L}$ is addressed by sensitivity analysis, evaluating the versions of Table 1 for alternative (plausible) loss matrices. If we reduce its entries L_{31} and L_{13} from 5 to 4, we obtain the same verdicts as in Table 1, even though the expected losses for the extreme verdicts 1 and 3 are reduced. Verdicts 1 and 3 become more attractive because the loss for issuing them inappropriately is less harsh. An increase of L_{31} and L_{13} , say, from 5 to 7, is appropriate for a client who is more averse to the verdicts of inequality. One verdict is then changed, from 1 to 2 for network 7.

A distinctly asymmetric loss matrix $\mathbf{L}$ is appropriate for a client who is liberal with praise for good performance (high rate θ_k), but reluctant to sanction a poor performance, or vice versa. The loss matrix

$$\begin{pmatrix} 0 & 1 & 10 \\ 3 & 0 & 1 \\ 5 & 3 & 0 \end{pmatrix}$$

represents such a perspective. With it, the verdicts for networks 11 and 14 are changed from 3 to 2, whereas the verdicts for networks 12 and 13 remain at 3. The verdicts are not aligned with the estimates $\hat{\theta}_k$, because of the influence of the sampling variances of estimators. Large variance increases the plausibility of higher values of θ_k ; indeed, the standard errors of $\hat{\theta}_{11}$ and $\hat{\theta}_{14}$ are much greater than the standard errors of $\hat{\theta}_{12}$ and $\hat{\theta}_{13}$. This can be interpreted as networks 11 and 14 being granted the ‘benefit’ of greater plausibility of high rates of compliance.

To underline the dependence of the verdicts on the loss matrix $\mathbf{L}$, consider the matrix

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 6 \\ 3 & 0 & 2 \\ 15 & 8 & 0 \end{pmatrix},$$

which reflects a strong aversion to overstatement as well as to confusion of the two extreme categories. The result is a ‘boost’ to category 2, at the expense of category 3. Now verdict 1 is issued to networks 1–7, verdict 2 to networks 8–14, and verdict 3 only to networks 15 and 16.

The inequality verdicts 1 and 3 become more attractive when the similarity range (T_1, T_2) is made narrower. Indeed, by setting $T_1 = T_2$ verdict 2 is eliminated altogether.

7.2 Exceptional networks

Figure 3 presents the plot of the exceptionality functions $P(\Lambda, G)$, given by (3), in the range $\Lambda \in (40, 75)\%$ and for values of the gap G between 2 and 6%, indicated on the curves. The curve for $G = 4\%$ is highlighted. The diagram confirms that a narrower gap is associated with uniformly greater probability of presence of exceptional observations. The functions attain their local maxima about halfway between consecutive observations that are further apart. Their local minima are at the observed rates, marked by black discs at the top and bottom of the diagram.

FIGURE 3

One of a set of local maxima is close to $\Lambda_1 = 45.3\%$ for all values of G ; $P(\Lambda_1; 0.02) = 0.771$ and $P(\Lambda_1; 0.03) = 0.659$. (The maxima are not exactly at the same value of Λ .) Since Λ_1 separates $\hat{\theta}_1$ from $\hat{\theta}_k$ for $k > 1$ by a wide margin, we conclude that network 1 is very likely to be exceptionally poor for $G \leq 2\%$. Similarly, based on the local maxima close to $\Lambda_{15} = 69.3\%$, where $P(\Lambda_{15}; 0.04) = 0.753$, we conclude that network 16 is very likely to be exceptionally good for $G \leq 4\%$. Of course, we have to qualify the term ‘very likely’; we defined it as $P > 0.75$.

Let R be the loss ratio for failing to declare a gap between consecutive ordered observations (e.g., at 69.3%) as a separator — separating exceptional networks in the tail of the distribution from the mainstream networks. Then choosing a separator is associated with expected loss $1 - P(\Lambda, G)$, owing to the possibility that the declaration is inappropriate (a false positive), and failing to choose it (a false negative) with expected loss $RP(\Lambda, G)$. This leads to the rule

$$P(\Lambda, G) > \frac{1}{R + 1}$$

(or $R > 1/P - 1$) for declaring a separator of width G at Λ . The loss ratio R may incorporate the number of networks at stake. More conservative attitudes correspond to lower values of R , and higher values of R reflect more liberal attitudes to identifying exceptional networks. The choice of $R = 1$ might appear as a good compromise. It should not be regarded as a default because it corresponds to a particular attitude or perspective, and would make the method too restrictive. Ambiguity about R can be

dealt with by sensitivity analysis — reporting the results for a range of plausible values of R . This is equivalent to reporting the results for the limits of the plausible range of R .

For $G = 4\%$, the exceptionality function $P(\Lambda, G)$ attains its local maximum of [0.548](#) at $\Lambda = 45.3\%$. This value, or the associated gap $\Lambda \pm \frac{1}{2}G$, separates network 1 from the rest, and network 1 is exceptionally poor for $R > 1/0.548 - 1 = \textcolor{blue}{0.825}$. The function $P(\Lambda, 0.04)$ attains another local (and the global) maximum of [0.756](#) at $\Lambda = 69.3\%$. It separates network 16 from the rest. The network is exceptionally good for $R > \textcolor{blue}{0.323}$. If the lower limit of the plausible range for R is $R_L > 0.323$, then network 16 has performed exceptionally well, and this verdict is unequivocal for $G \leq 4\%$. If $R_L > 0.825$, then we have an unequivocal verdict that network 1 has performed exceptionally poorly.

7.3 Winners and losers

Figure 3 suggests that identifying the networks with the highest and lowest rates of compliance entails relatively little uncertainty. This is confirmed by evaluating the probabilities of the corresponding two events. These probabilities are 0.997 for network 16 and 0.969 for network 1. The probability of network 2 being a loser is 0.017, so it would be singled out together with network 1 only for $R > 57.3$. Network 14 has a higher probability of being a winner than network 15, 0.0012 vs. 0.0008. This is not a contradiction with the inequality $\hat{\theta}_{14} < \hat{\theta}_{15}$, but a consequence of the disparity of the standard errors of the two estimators. The probability of the event $\theta_{14} > 0.70$ is very small, but nevertheless greater than for θ_{15} , even though $\hat{\theta}_{14} < \hat{\theta}_{15}$.

If we introduce the hurdle for a winner that it must exceed 75%, then the probability of network 16 being a winner is only 0.298; the pseudo-network 17 is the winner with probability 0.702. If we intended to sanction the loser only if its rate of compliance is below 45%, then network 1 would still be sanctioned. The probability that the pseudo-network with $\theta_{17} = 0.45$ is a loser is smaller than 0.001, and the high probability for network 1 is unaffected.

7.4 Classification

Suppose classes are defined by the cut-points at 50, 60 and 70%, and the loss is defined as the number of cut-points missed, $L_{vV} = |v - V|$. Then network 1 is assigned to class 1 ($\theta_1 < 50\%$), networks 2–10 to class 2 (50–60%), networks 11–15 to class 3 (60–70%) and network 16 to class 4 ($> 70\%$). There may be some contention about the classification of two networks, 2 and 10, because the respective classes 1 and 3 are close contenders for them. For network 2, $Q_1 = 0.545$ and $Q_2 = 0.455$, and for network 10, $Q_2 = 0.489$ and $Q_3 = 0.511$. The classification of these two networks is altered by introducing a slight asymmetry in the loss. Preferring the error of overstating the class (that is, aversion to understatement) is introduced by a loss ratio $R < 1$. The verdict for network 10 is changed from 2 to 3 for $R \leq 0.95$. Aversion to overstatement is introduced by setting $R > 1$. For $R \geq 1.2$, the verdict for network 2 is changed from 2 to 1.

Loss ratio $R < 1$ is appropriate if the audit is keener to highlight good performance than to identify poor performance. In contrast, $R > 1$ is appropriate if the focus of the audit is on identifying poor practice. Even within a single industry or service, such as health care, the foci of the various audits are different and cannot be characterised by a single loss ratio. In areas where the scope for improvement is relatively easy to pinpoint, and where most units lag behind a standard or a goal, the audit may be liberal with naming institutions as poor performers and maintain a high bar for any rewards for outstanding performance. In other settings, liberal distribution of rewards may have a positive effect on the morale of staff and management and may encourage cooperation in initiatives aimed at improving the service. Without a clearly one-sided objective of the audit, $R > 2$ and $R < 0.5$ are unlikely to be appropriate, and (0.5, 2) or its subinterval can be used as a plausible range for R .

The error of crossing one cut-point may be associated with greater loss than the errors of crossing other cut-points. Also, the errors of crossing cut-points need not be additive. For example, L_{31} may be set to a value greater than $L_{32} + L_{21}$. These features of loss are easy to incorporate by specifying the loss matrix $\mathbf{L}$ directly.

7.5 Being one of the best

The top panel of Figure 4 presents the probabilities $\max_{\Lambda} W_k(\Lambda, G)$ of network k being one of the best for a range of tolerance gaps G . It confirms that network 16 has a peer in network 15 only for a very wide gap G and a very liberal assessment, characterised by a small value of the loss ratio R . For $G = 10\%$, $\max_{\Lambda} W_{15} = 0.326$ and $\max_{\Lambda} W_{14} = 0.170$, so network 15 would be declared as one of the best only by a very liberal criterion even with such a wide gap. Network 14 is included only with an extremely liberal criterion.

FIGURE 4

For the poorest performing networks we have similar verdicts. Networks 2 and 3 join network 1 as one of the worst only for very wide gaps and liberal criteria.

8 Comparing hospitals

Although the focus of our study is on networks, we illustrate the methods on comparing hospitals as an example of dealing with many units. From the 177 hospitals in the data we exclude 25 that have fewer than 10 infants each in the studied population. Their assessment is associated with too much uncertainty.

The assumption of normality is problematic for several hospitals with small caseloads and the highest and lowest observed rates $\hat{\theta}_k$ because we have to reckon with the boundaries of 0 and 1 for θ_k . [This is addressed by applying a beta empirical prior.](#) (The beta distribution is conjugate for the binomial distribution.) Suppose there are m_k instances of compliance among n_k infants, and let the prior for the underlying rate θ_k be beta with parameters α and β . Then the posterior distribution of θ_k is also beta, with parameters $m_k + \alpha$ and $n_k - m_k + \beta$. The prior parameters α and β can be interpreted as additional (prior) numbers of cases of compliance and failures. [In an empirical Bayes approach, \$\alpha\$ and \$\beta\$ are estimated by maximum likelihood.](#)

All the formulae derived in Sections 2–6 are adapted straightforwardly by replacing each value of the normal probability with the probability of the appropriate beta distribution. The only exception is for comparing the rates of a hospital and the country in equation (6), where the (small) uncertainty about the national rate θ could be dealt with by integrating the conditional expected losses given the value of θ over its

posterior distribution. This is not necessary because even the largest caseload of the hospitals, $\max_k n_k = 162$, is a small fraction of the overall caseload $n = 7783$.

Figure 5 presents the caterpillar plot of the hospital-level rates of compliance with the standard for first measurement of temperature. It incorporates the comparisons of each hospital with the national rate for the thresholds $T_1 = -5\%$ and $T_2 = 5\%$, and the loss matrix given in (7). The estimates for hospitals are marked by black, grey and white discs for the respective verdicts 1 ($\theta_k \prec \theta$), 2 ($\theta_k \simeq \theta$) and 3 ($\theta_k \succ \theta$). The respective verdicts 1–3 are issued to 50, 33 and 69 hospitals. These hospitals do not have a contiguous set of ranks because the sampling variation of their estimators is a non-trivial factor — for large hospitals (with small σ_k^2) verdict 2 is more attractive. The verdict for hospitals 82 and 85 is 3, whereas for hospitals 83 and 84 it is 2.

FIGURE 5

The numbers of hospitals in the categories 1–3 (i.e., the sizes of the categories) are not necessarily well estimated by counting the verdicts. The estimator with minimum mean squared error is defined by the totals of the posterior probabilities of belonging to the categories. The mean squared error may not represent the client’s perspective well. In an analysis that adheres more closely to our general proposal, a loss function is declared and each count is estimated by minimising the expected loss.

For classifying hospitals into categories defined by the cut-points 40, 50, $\dots$, 70%, we consider the loss ratios in the range (0.8, 1.2). The verdicts are unequivocal (identical for all plausible values of R) for 11, 19, 31, 37 and 28 hospitals, 126 in total, to respective categories 1–5. For each of the remaining 26 hospitals, the verdicts are in two adjacent categories, for 2, 4, 12 and 8 hospitals in the respective pairs of categories 1–2, $\dots$, 4–5. In all instances, the verdict switches to a lower category as R is increased.

The probabilities of separation of units with exceptionally low rates of compliance are plotted in Figure 6 for the gaps of 1–5%. The diagram shows that the three units with estimated rates lower than 30% are likely to be exceptional for relatively narrow gaps and a liberal perspective. Unlike in the corresponding Figure 3 for networks, the functions $P(\Lambda, G)$ are quite smooth, with only a slight ‘bump’ between the estimates $\hat{\theta}_3$ and $\hat{\theta}_4$ which are wide apart. This is a consequence of the large sampling variation

of these estimators — the relative distances of the estimates θ_k are a poor indication of the distances of the targets θ_k .

FIGURE 6

Figure 7 displays the probabilities W_k of being nearly as good as the winner, as functions of the tolerance gap G . These probabilities are very small for most hospitals; a few hospitals with the highest probabilities are indicated on their curves. Even for gap $G = 10\%$ and the hospitals with the largest values of $\hat{\theta}_k$, this probability is smaller than 0.5. This is a consequence of the sizeable sampling variation of the rates — a large value of $\hat{\theta}_k$ does not preclude having a value θ_k much smaller than the maximum.

FIGURE 7

9 Discussion

In everyday life, we make numerous decisions, both momentary and after carefully weighing the possible advantages and drawbacks of the contemplated courses of action, together with the uncertainties involved. In many instances, a short list of options is available. We choose one of the options not solely based on their likelihood of being the best, but consider also the repercussions (ramifications) should the choice not be optimal.

We prefer to integrate these two statistical activities, inference and selection (decision), into a single evaluation that concludes with the preferred course of action. The analysis is not objective (universal or multipurpose) but is tailored to the particular perspective of the client or the sponsor of the analysis. The approach requires a close cooperation of the analyst with the client to quantify this perspective. Established approaches based on hypothesis testing are objective, but they correspond to a particular (default) perspective that may be distant from the client's perspective.

The loss (matrix), thresholds, cut-points and related parameters represent the client's perspective. They are essential inputs in our approach. The approach is open to abuse if the client, possibly in cooperation with the analyst, attempts to second-guess the inputs for which an a priori choice (a rigged verdict) would be obtained. In this case, the analysis serves a false purpose because the perspective is not formulated

independently of the observed data. Similar problems arise when the analyst communicates with the client indirectly through an intermediary. For example, the client may be a party, such as the general public, with whom direct communication is not feasible. An authority charged with representing the public may misjudge the public's priorities or deliberately misrepresent them, and further biases or errors may be introduced in the process of elicitation. An organisation may appoint a representative to deal with the analyst; his or her good faith and accurate communication of the organisation's perspective are imperative for the validity of the analysis. These threats to validity of the method should not discourage the analyst because the alternative is an analysis with a default perspective which may not be relevant.

Interpretation of the results of an analysis concluding with verdicts of hypothesis tests, estimates (with the associated standard errors), confidence intervals, or the like, is subject to similar threats, except that even an honest client has to face insurmountable difficulties in converting them to effective action without familiarity with statistical evaluations and intimate understanding of uncertainty.

The methods [presented](#) in Sections 2–6 are based on normally distributed posterior distributions and piecewise constant losses. Extensions to other common distributions, such as Student t, beta (see Section 8) and gamma, are straightforward because the evaluations do not use any properties that are specific to the normal distribution. Several examples are analysed in [7]. The loss may depend not only on the kind of the error committed (e.g., false positive or false negative), but also on the magnitude of the error. This perspective is catered for by piecewise linear, quadratic and other loss functions. The loss structures can be defined separately for the distinct kinds of error. Further flexibility is afforded by linearly combining loss functions. The key property of a loss function is linearity, that it can be treated like a (monetary) currency for error and that it faithfully reflects our intent to be frugal in its expenditure on the process of making well informed decisions.

Hierarchical (two-level) models are commonly applied to performance assessment, [21] and [22], with the two levels represented by elements (patients) and aggregate-level units (networks), or three levels (elements, hospitals and networks). These models assume that the aggregate units are a random sample from a (super-)population. Our objection to these models in the frequentist paradigm is based on considering hypo-

thetical replications of the study. Since the assessment is for specific networks (or hospitals), a replication of the study would involve a different set of elements (patients), but the same set of hospitals and networks. Thus, hospitals and networks are *fixed* units of analysis. In contrast, in hierarchical models these aggregate-level units are random, so the replication scheme would involve a different set of aggregate-level units.

Our main objection is not to using estimators based on hierarchical models, but to claims and, more generally, assessments of estimators, that are based on the assumptions associated with these models. For example, the (empirical Bayes) shrinkage estimators are regarded by many as unbiased and efficient. In Appendix we show that this is not the case. In particular, they are problematic for units that are extreme — that have the smallest or largest values. In some assessments, these units are the focus of the analysis. In the spirit of Box [23], we do not object to applying models that are invalid, but to assessment of the estimators they yield on the terms of these models. The assessment should be based on a model that is more realistic. In our case, it entails evaluations based on the assumption that θ_k are fixed. The maximum likelihood estimator based on this assumption is dismissed not because it is invalid, but because it is inefficient, involving too many parameters (degrees of freedom).

The additional (extra-data) input that our approach requires introduces a complexity and in some settings also contention. However, it is essential for tailoring the analysis to the client’s purpose and enables us to move a step closer to making informed and practically relevant decisions. Difficulties with quantifying the client’s perspective are addressed by using plausible ranges of the relevant parameters, which can be regarded as a form of sensitivity analysis. The approach we propose integrates statistical analysis, as currently perceived, with decision making, which is often left unattended or left for the client to contend with.

Acknowledgements

Discussions with several members of the NNAP Project Board on the subject treated in this paper are acknowledged. The opinions expressed in this paper are solely the author’s and do not reflect the policies of the Board nor any decisions made collectively by the Board. All the perspectives (loss ratios and matrices) considered in the paper

are the author’s guesses, set without any elicitation from the Board or other experts. I would like to thank Daniel Gray for extracting the relevant data from NNRD and for assisting me with orientation in the database.

References

- [1] Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, Blunt I, Wood C, Grigg O. Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society Series A* 2012; **175**:1–47.
- [2] Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statistics in Medicine* 2005; **24**:1185–1202.
- [3] Goldstein H, Healy, MJR. The graphical presentation of a collection of means. *Journal of the Royal Statistical Society Series A* 1995; **158**:175–177.
Graphical methods *Journal of the Royal Statistical Society*
- [4] Longford NT. Decision theory applied to selecting the winners, ranking, and classification. *Journal of Educational and Behavioral Statistics* 2016; **41**:420–442.
- [5] Berger JO. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer-Verlag: New York, 1985.
- [6] Lindley DV. *Making Decisions*. 2nd ed. Wiley: Chichester, UK, 1985.
- [7] Longford NT. *Statistical Decision Theory*. Springer-Verlag: Heidelberg, Germany, 2013.
- [8] Gelman A, Stern H. The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician* 2006; **60**:328–331.
- [9] Longford NT. Comparing two treatments by decision theory. *Pharmaceutical Statistics* 2016; **15**:387–395.
- [10] Longford NT. Variation of the rates of necrotising enterocolitis in the neonatal networks in England. *Statistics in Medicine* 2017; **36**:1655–1668.

- [11] Fisher RA. Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B* 1955; **17**:69–78.
- [12] Zabell SL. R.A. Fisher and the fiducial argument. *Statistical Science* 1992; **7**:369–387.
- [13] Hannig J. On generalized fiducial inference. *Statistica Sinica* 2009; **19**:491–544.
- [14] Lindley DV. Fiducial distribution and Bayes’ theorem. *Journal of the Royal Statistical Society Series B* 1958; **20**:102–107.
- [15] Shafer JP. Multiple hypothesis testing: a review. *Annual Review of Psychology* 1995; **46**:561–584.
- [16] Benjamini Y, Yekutieli D. The control of the false discovery rate under dependency. *Annals of Statistics* 2001; **29**:1165–1188.
- [17] Seaton SE, Barker L, Lingsma HF, Steyerberg EW, Manktelow BN. What is the probability of detecting poorly performing hospitals using funnel plots? *British Medical Journal, Quality and Safety* 2013.
- [18] National Clinical Audit Advisory Group. *Detection and management of outliers*. Department of Health: London, 2011.
- [19] Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 2002.
- [20] van Buuren S. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC: Boca Raton, FL, 2012.
- [21] Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. *Statistics in Medicine* 1994; **13**:889–903.
- [22] Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Series A* 1996; **159**:385–443.
- [23] Box GEP. Science and statistics. *Journal of the American Statistical Association* 1976; **71**:791–799.

Appendix

Suppose in addition to the assumptions about the sampling distributions $(\hat{\theta}_k | \theta_k) \sim \mathcal{N}(\theta_k, \sigma_k^2)$ we also have that θ_k , $k = 1, \dots, K$, are a random sample from $\mathcal{N}(\theta, \tau^2)$. For simplicity, we assume that θ and τ^2 are known. The empirical Bayes (shrinkage) estimator of θ_k is

$$\tilde{\theta}_k = (1 - b_k) \hat{\theta}_k + b_k \theta,$$

where $b_k = \sigma_k^2 / (\sigma_k^2 + \tau^2)$. In fact, $\tilde{\theta}_k$ is the conditional expectation of θ_k given the data and the values of the parameters. As a substitute for its sampling variance, the conditional variance of θ_k ,

$$\text{var}(\theta_k | \hat{\theta}_k; \theta, \tau^2) = \frac{\sigma_k^2 \tau^2}{\sigma_k^2 + \tau^2}, \quad (8)$$

is commonly used.

If θ_k is regarded as fixed, then the bias of $\tilde{\theta}_k$ is $B_k = -b_k(\theta_k - \theta)$ and its sampling variance is $V_k = (1 - b_k)^2 \sigma_k^2$. Therefore its mean squared error is

$$\begin{aligned} \text{MSE}(\hat{\theta}_k; \theta_k) &= V_k + B_k^2 \\ &= \frac{\sigma_k^2 \tau^2}{(\sigma_k^2 + \tau^2)^2} \left\{ \tau^2 + \sigma_k^2 \frac{(\theta_k - \theta)^2}{\tau^2} \right\}. \end{aligned}$$

This is equal to the expression in (8) only when $\theta_k = \theta \pm \tau$, that is, when θ_k is in the ‘typical’ distance from θ . The MSE exceeds $\text{var}(\hat{\theta}_k) = \sigma_k^2$ when

$$(\theta_k - \theta)^2 > 2\tau^2 + \sigma_k^2.$$

This occurs for a few units k with the highest and lowest values of θ_k . **They are not necessarily outliers.** For example, if $\sigma_k^2 = \frac{1}{2}\tau^2$, this condition is equivalent to $\theta_k > \theta + 1.58\tau$ or $\theta_k < \theta - 1.58\tau$, and this is satisfied in a typical dataset by about 11.5% of units. In conclusion, the empirical Bayes estimator is problematic for extreme units which happen to be the most important in a typical analysis.

Table 1: Expected losses for the comparisons of the networks' rates with the national rate for the standard of first measurement of body temperature after birth. The thresholds are $T_1 = -2.5\%$ and $T_2 = 2.5\%$ and the loss matrix is given in (7).

	Network							
	1	2	3	4	5	6	7	8
Q_1	<u>0.000</u>	<u>0.024</u>	<u>0.038</u>	<u>0.101</u>	<u>0.210</u>	<u>0.257</u>	<u>0.849</u>	0.957
Q_2	2.000	1.953	1.924	1.798	1.613	1.501	0.944	<u>0.490</u>
Q_3	5.000	4.905	4.848	4.596	4.210	3.995	2.567	1.778
	9	10	11	12	13	14	15	16
Q_1	1.046	2.089	2.946	3.119	4.035	3.942	4.937	5.000
Q_2	<u>0.539</u>	<u>0.548</u>	1.055	1.071	1.519	1.489	1.969	2.000
Q_3	1.762	0.733	<u>0.637</u>	<u>0.488</u>	<u>0.244</u>	<u>0.291</u>	<u>0.016</u>	<u>0.000</u>

Note: The labels of the networks, 1 – 16, are the same as in Figure 2.

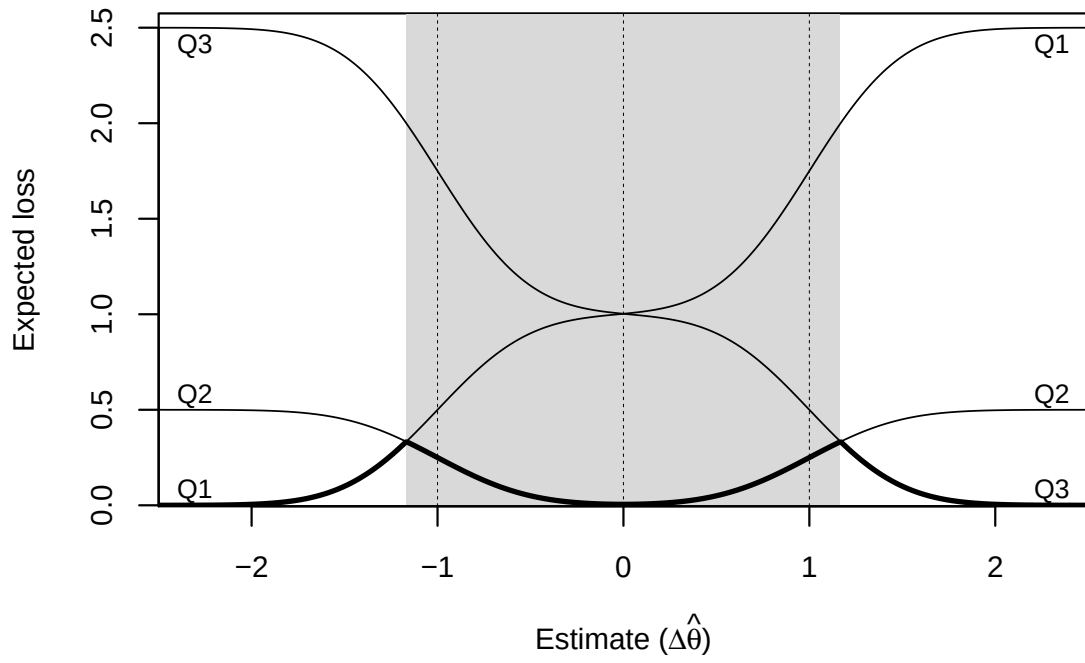


Figure 1: Expected losses as functions of the estimate $\Delta\hat{\theta}$ (standard error $\Delta\sigma = 0.39$) and loss matrix given in (1) for comparing two means with thresholds ± 1 .

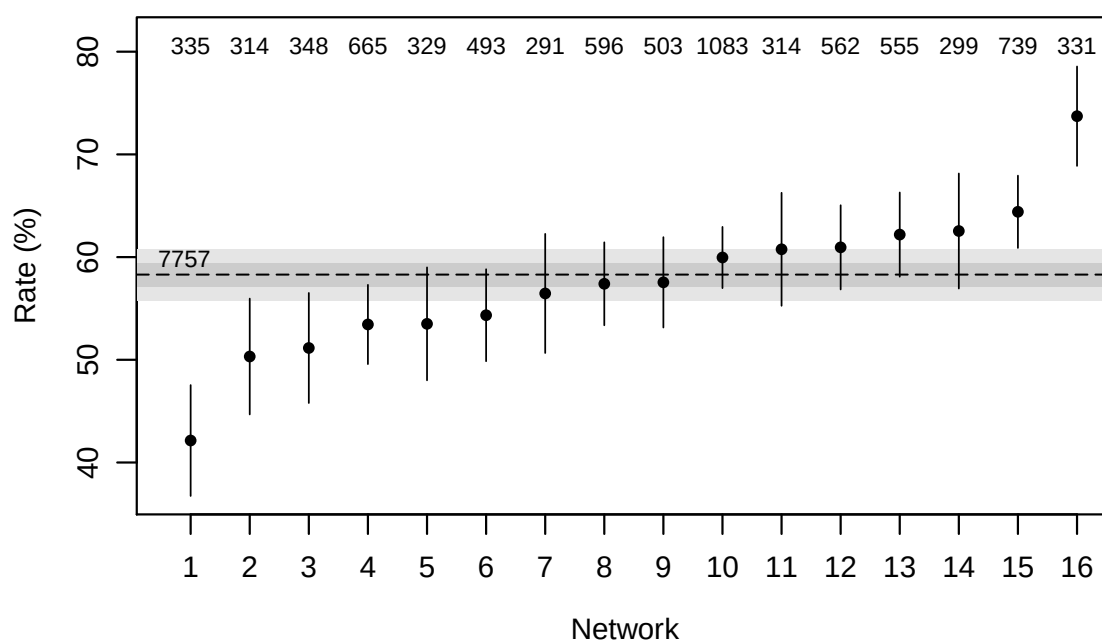


Figure 2: Caterpillar plot of the network-level rates of compliance with the standard for the first measurement of body temperature. The caseloads (numbers of qualifying infants in care) for the networks are printed at the top margin. The national caseload is printed at the horizontal dashes that mark the national rate. The grey horizontal strips are explained in the text.

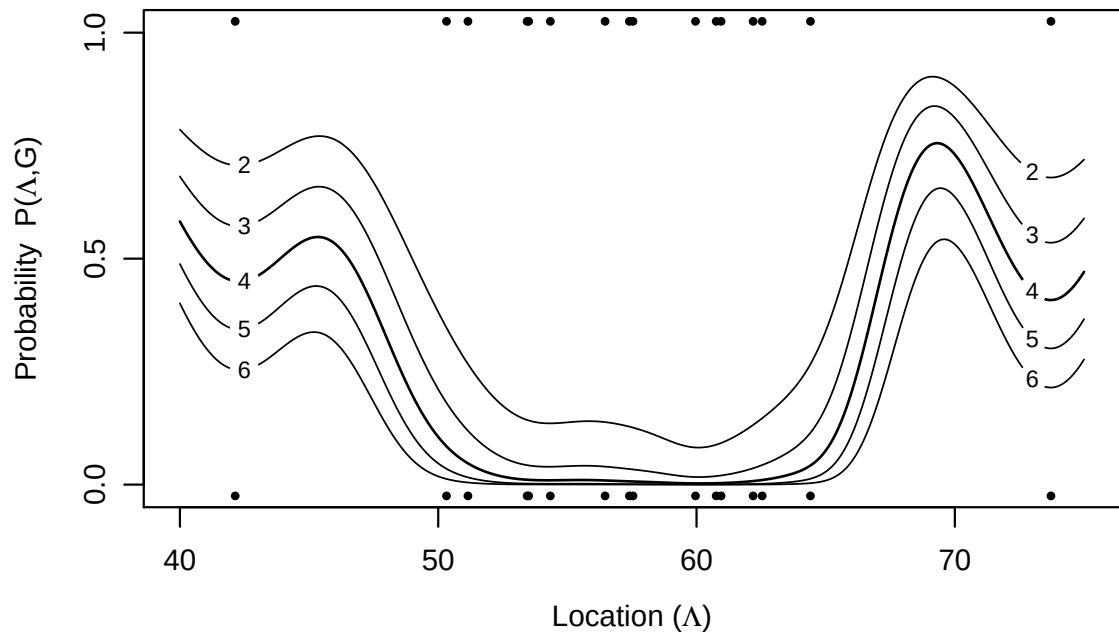


Figure 3: Exceptionality plot for the network-level rates of compliance with the standard for the first measurement of body temperature.

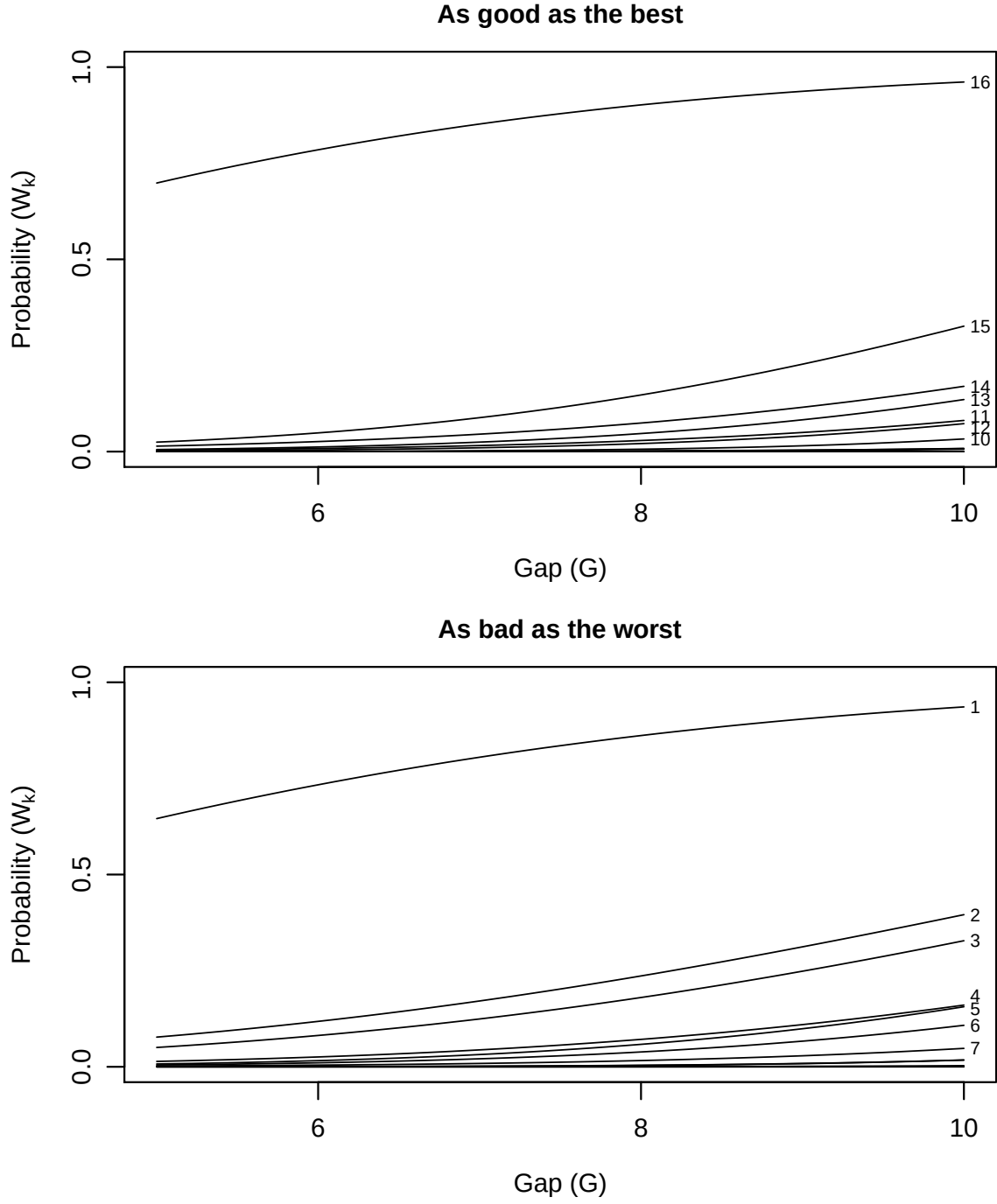


Figure 4: Probabilities $\max_{\Lambda} W_k(\Lambda, G)$ as functions of the gap G . Some of the networks are identified at the right-hand margins.

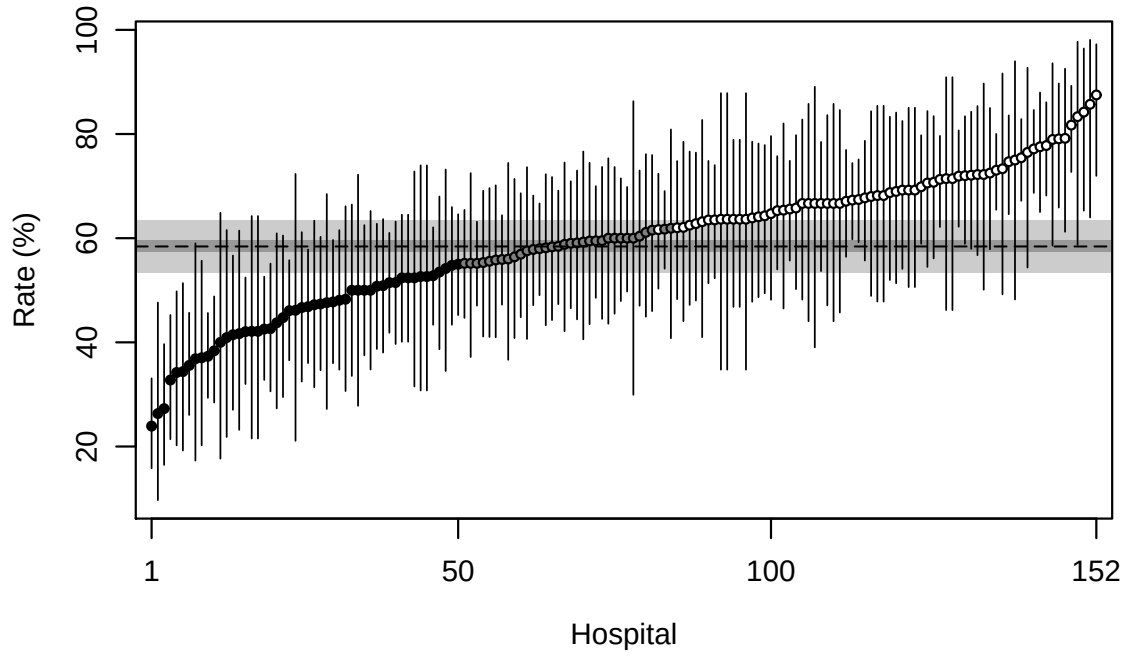


Figure 5: Caterpillar plot of the hospital-level rates of compliance. The estimates $\hat{\theta}_k$ are marked according to the verdict by black (verdict 1), grey (2) or white discs (3).

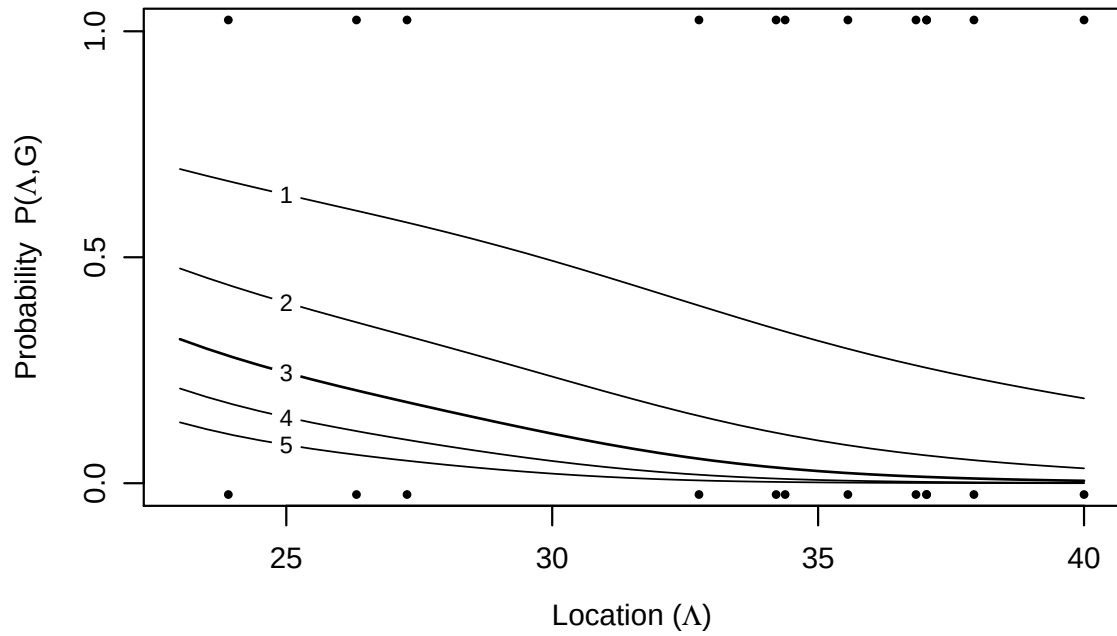


Figure 6: Exceptionality plot for the hospital-level rates of compliance with the standard for body temperature.

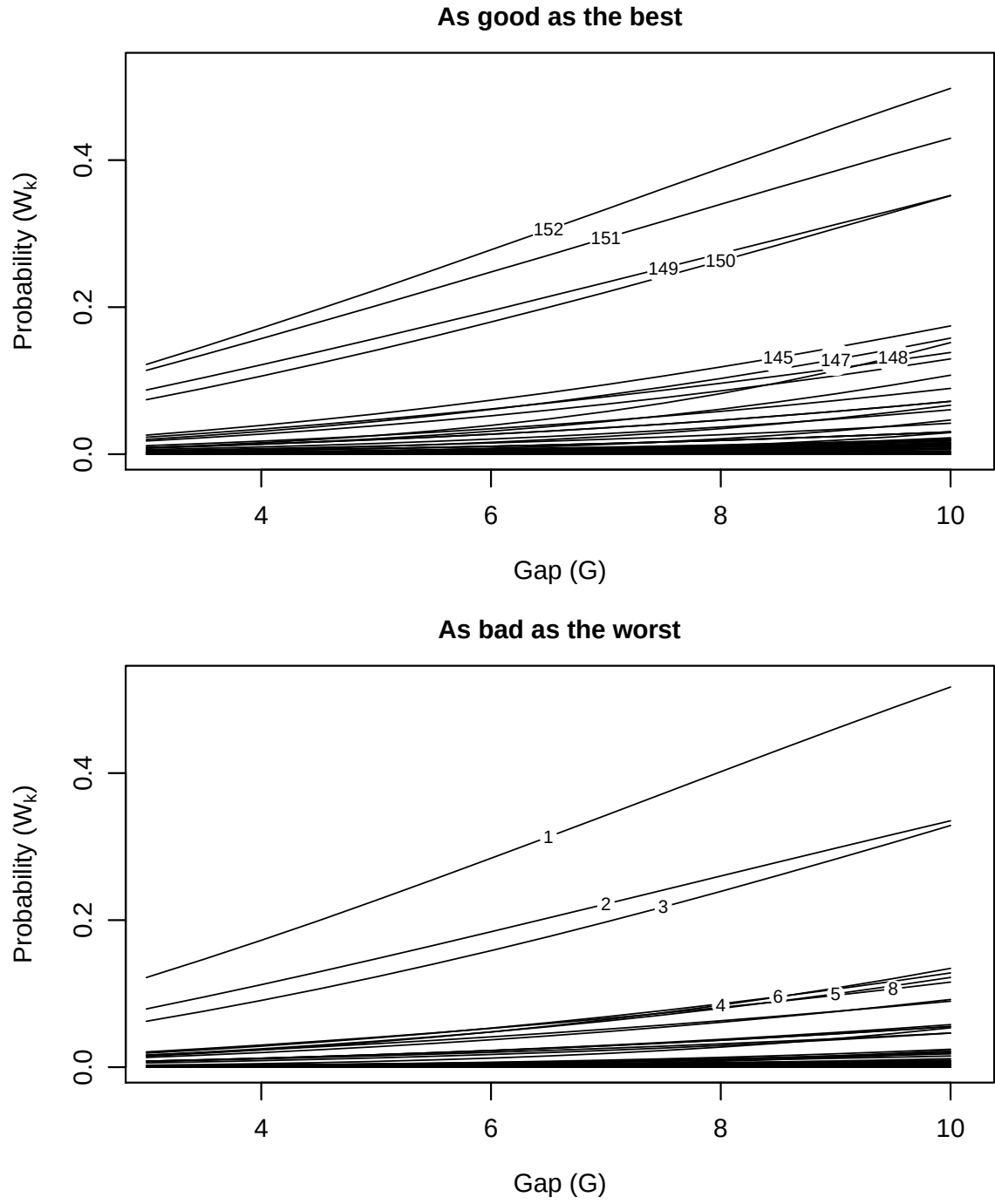


Figure 7: Hospital-level probabilities $\max_{\Lambda} W_k(\Lambda, G)$ as functions of the gap G . Some of the hospitals are identified on the curves.