Ranking USRDS provider-specific SMRs from 1998–2001

Rongheng Lin,[1,*] **Thomas A. Louis,**[2]
Susan M. Paddock[3] **and Greg Ridgeway**[3]

[1] Department of Biostatistics, Johns Hopkins University
615 N.Wolfe Street, Baltimore, MD 21205, U.S.A.
Phone: 410-614-5086 FAX: 410-955-0958
[2] Department of Biostatistics, Johns Hopkins University
[3]Rand Statistics Group, Santa Monica, CA 90407 U.S.A.

December 14, 2004

Supported by grant 1-R01-DK61662 from U.S NIH
National Institute of Diabetes, Digestive and Kidney Diseases

---

[*] *email:* rlin@jhsph.edu

Summary

Provider profiling (ranking, "league tables") is prevalent in health services research. Similarly, comparing educational institutions and identifying differentially expressed genes depend on ranking. Effective ranking procedures must be structured by a hierarchical (Bayesian) model and guided by a ranking-specific loss function, however even optimal methods can perform poorly and estimates must be accompanied by uncertainty assessments. We use the 1998-2001 Standardized Mortality Ratio (SMR) data from United States Renal Data System (USRDS) as a platform to identify issues and approaches. Our analyses extend Liu et al. (2004) by combining evidence over multiple years via an AR(1) model; by considering estimates that minimize errors in classifying providers above or below a percentile cutpoint in addition to those that minimize rank-based, squared-error loss; by considering ranks based on the posterior probability that a provider's SMR exceeds a threshold; by comparing these ranks to those produced by ranking MLEs and ranking P-values associated with testing whether a provider's $SMR = 1$; by comparing results for a parametric and a non-parametric prior; by reporting on a suite of uncertainty measures.

Results show that MLE-based and hypothesis test based ranks are far from optimal, that uncertainty measures effectively calibrate performance; that in the USRDS context ranks based on single-year data perform poorly, but that performance improves substantially when using the AR(1) model; that ranks based on posterior probabilities of exceeding a properly chosen SMR threshold are essentially identical to those produced by minimizing classification loss. These findings highlight areas requiring additional research and the need to educate stakeholders on the uses and abuses of ranks; on their proper role in science and policy; on the absolute necessity of accompanying estimated ranks with uncertainty assessments and ensuring that these uncertainties influence decisions.

KEY WORDS: Ranks/percentiles; Bayesian hierarchical model, Uncertainty assessments

1. Introduction

Performance evaluations of health services providers burgeons (Christiansen and Morris (1997), Goldstein and Spiegelhalter (1996), Landrum et al. (2000), Liu et al. (2004), McClellan and Staiger (1999)). Similarly, analyzing spatially structured health information (Conlon and Louis (1999), Louis and Shen (1999), Shen and Louis (2000)), ranking teachers and schools (Lockwood et al. (2002)), identification of differentially expressed genes (Newton et al. (2001), Dudoit et al. (2002), Kendziorski et al. (2004)) are increasing in prevalence and importance. Goals of such investigations include valid and efficient estimation of population parameters such as the average performance (over clinics, physicians, health service regions or other "units of analysis"), estimation of between-unit variation (variance components), and unit-specific evaluations. These latter include estimating unit specific attributes, ranking units for profiling and league tables (Goldstein and Spiegelhalter (1996)), identification of excellent and poor performers, the most differentially expressed genes, and determining "exceedences" (how many and which unit-specific true parameters exceed a threshold).

We present analysis of the information in the U. S. Renal Data System (USRDS) as a platform for our approaches. The Standardized Mortality Ratio (SMR), the ratio of observed to expected deaths, is an important service quality indicator (see Zaslavsky (2001)). The USRDS produces annual estimated SMRs for several thousand dialysis centers and uses these as a quality screen (Lacson et al. (2001), ESRD (2000), USRDS (2003)). Invalid estimation or inappropriate interpretation can have serious local and national consequences.

Though estimating the SMRs is a standard statistical operation (produce provider-specific expected deaths based on a statistical model, and then compute the "observed/expected" ratio), dealing with complications such as the need to specify a reference population (providers included, the time period covered, attribution of events); the need to validate the model used to adjust for important patient attributes (age, gender, diabetes, type of dialysis, severity of disease); the

need to adjust for potential biases induced when attributing deaths to providers, and accounting for informative censoring are important and challenging. From the USRDS we obtained observed and expected deaths for the $K = 3173$ dialysis centers that contributed information for all four years. We accept the USRDS approach for producing these values and focus on using them as evidence on SMRs.

The data structure and inferential goals require a hierarchical, random effects model that accounts for nesting relations and specifies both population values and random effects. Correctly specified, a hierarchical model properly accounts for the sample design and provides the necessary structure for developing scientific and policy-relevant inferences. Bayesian hierarchical models effectively accomplish these goals, accounting for variance components and other uncertainties, improving estimation of random effects. Valid and efficient estimation of population parameters, variance components, and unit-specific random effects (provider-, clinician-, region- or gene-specific latent attributes), all in the context of valid uncertainty assessments.

As Shen and Louis (1998) show and Gelman and Price (1999) present in detail, no single set of estimates or assessments can effectively address multiple goals and we provide a suite of assessments. Guided by a loss function, the approach structures non-standard inferences such as ranking (including identification of extremely poor and good performers) and estimating the histogram of random unit-specific effects. As shown by Liu et al. (2004), when estimation uncertainty varies over units, ranks produced from hypothesis test statistics evaluating whether a unit's SMR $= 1$ inappropriately identify units with relatively low variance as extreme because these tests have highest power; ranks produced from the MLEs inappropriately identify units with relatively high variance as extreme. Effective ranking depends on properly accommodating both signal and noise.

We present a variety of ways to rank/percentile SMRs (or other unit-specific attributes) and

to measure performance of the comparisons. These comparisons are relative, and our focus on them in no way implies that estimation of the attribute itself (e.g., provider-specific SMR or death rate) is of less relevance. Our work extends Liu et al. (2004) by considering new ranking/percentiling methods (see Lin et al. (2004)) and combining evidence over multiple years of data. We develop a hierarchical model that allows for a first-order, autoregressive (AR1) structure relating the log(SMRs) for each provider over the four years and compare estimated percentiles based on single-year and multiple-year analyses, the latter "borrowing information" across years. We apply and compare several ranking/percentiling methods, including those that minimize Squared Error Loss (SEL) for ranks (see Shen and Louis (1998)), ranks/percentiles based on optimal classification, those based on ranking exceedance probabilities (see Normand et al. (1997)) and those produced by ranking either the MLE estimates of unit-specific attributes or unit-specific statistics testing whether a unit's attribute differs from the typical value. Multiple year analyses are based on a log-normal prior and for the single-year analyses, we implement and compare results based on it and on the Non-parametric, maximum likelihood (NPML) estimated prior.

Ranking and otherwise comparing is very uncertain and is based on both explicit and implicit assumptions (Liu et al. (2004), Lockwood et al. (2002)). Therefore, we implement a variety of comparisons and a variety of summaries are needed to communicate the full situation.

Section 2 presents our models; section 3 outlines several ranking methods; section 4 gives uncertainty measures; section 5 presents results and section 6 sums up and identifies additional research. Computer code for all routines is available at webpage http://www.biostat.jhsph.edu/~rlin/usrds-ranking.html.

2. Models

Let (Y_{kt}, m_{kt}) be the observed and expected deaths for provider k in year t, $t = 0, 1, 2, 3$ and $k = 1, \ldots 3173$ with the m_{kt} computed assuming that all providers give the same quality of care for patients with identical covariates. Let, $\rho_{kt} = \frac{E(Y_{kt})}{m_{kt}}$ so that $\rho_{kt} = 1$, if the provider has "average" performance. Let $\theta_{kt} = \log(\rho_{kt})$. We use a hierarchical, Poisson model with either a Gaussian prior (single-year and time series) for the θ_{kt} or a NPML prior for a single year. For the Gaussian prior, let $-1 < \phi < 1$ and use,

$$\begin{aligned} \xi_t \; iid \; N(0, V), \quad \phi \sim h_\phi(\cdot), \quad \lambda_t = \tau_t^{-2} \; iid \; \text{Gamma}(\alpha, \mu/\alpha) \qquad (1) \\ [\theta_{10}, \ldots, \theta_{K0} \mid \xi_0, \tau_0] \qquad iid \qquad N(\xi_0, \tau_0^2) \\ [\theta_{kt} | \theta_{k(t-1)}, \xi, \tau, \phi] \qquad ind \qquad N\left(\xi_t + \phi\tau_t\tau_{t-1}^{-1}\{\theta_{k(t-1)} - \xi_{t-1}\}, \{1-\phi^2\}\tau_t^2\right) \\ [Y_{kt} \mid m_{kt}, \rho_{kt}] \qquad \sim \qquad \text{Poisson}(m_{kt}\rho_{kt}). \end{aligned}$$

In all analyses based on (1) we use $V = 10, \mu = 0.01, \alpha = 0.05$ which stabilize the simulation while allowing sufficient flexibility. Single-year analyses result from setting $\phi \equiv 0$ in (1). For multi-year analyses via the AR(1) model, h_ϕ comes from using a N(0, 0.2) distribution for the Fisher's-z transformed ϕ, $z = 0.5\log\{(1+\phi)/(1-\phi)\}$. We implement the Gibbs sampler for model (1) with WinBUGS through the R package *R2WinBUGS*. The *coda* package was used to diagnose convergence.

For the single-year, NPML prior use,

$$[\rho_{1t}, \ldots, \rho_{Kt}] \; iid \; G; \;\; [Y_{kt} \mid m_{kt}, \rho_{kt}] \sim \text{Poisson}(m_{kt}\rho_{kt}), \qquad (2)$$

with G the NPML prior. See the appendix for the EM algorithm that computes the NPML and Laird (1978) and Carlin and Louis (2000) for additional information on it.

3. Ranking methods

A wide variety of methods are available for ranking including those based on ordering unit-specific summaries and those that operate through the posterior distribution of the ranks. The

former include ranking MLEs, posterior means (PMs) of target parameters, statistics for testing $H_0 : SMR_k = 1$ and exceedance probabilities (see below). The latter are produced by estimation of ranks via a loss function operating on them. If the posterior distributions of the unit-specific attribute are stochastically ordered, all (reasonable) approaches will agree.

We first define ranks and then specify candidate ranking methods. The ranks/percentiles of the ρ_{kt} are random variables induced by models (1) or (2). For clarity in defining ranks, we drop the index t and write $R_k(\boldsymbol{\rho}) = \text{rank}(\rho_k) = \sum_{j=1}^{K} I_{\{\theta_k \geq \theta_j\}}$ with the smallest ρ_k having rank 1. Ranks that depend only on the posterior distribution of the R_k give the same results for all monotone transforms of the target parameter (are monotone transform invariant).

3.1 Squared-error loss

Shen and Louis (1998) and Lockwood et al. (2002) study ranks that minimize squared error loss (SEL): $K^{-1}\sum_k (R_k^{est} - R_k(\boldsymbol{\rho}))^2$. It is minimized by the posterior expected ranks, $\bar{R}_k(\mathbf{Y}) = E_G[R_k(\boldsymbol{\rho}) \mid \mathbf{Y}] = \sum_{j=1}^{K} P_G[\rho_k \geq \rho_j \mid \mathbf{Y}]$, producing the optimal Mean Squared Error (MSE) equal to the average posterior variance of the ranks. Generally, the $\bar{R}_k$ are not integers; for optimal, distinct integer ranks, use $\hat{R}_k(\mathbf{Y}) = \text{rank}(\bar{R}_k(\mathbf{Y}))$.

In what follows, generally we drop dependency on $\boldsymbol{\rho}$ (equivalently, on $\boldsymbol{\theta}$) and omit conditioning on $\mathbf{Y}$. For example, R_k stands for $R_k(\boldsymbol{\theta})$ and $\hat{R}_k$ stands for $\hat{R}_k(\mathbf{Y})$. We present either ranks (R_k) or, equivalently, percentiles [$P_k = R_k/(K+1)$] with percentiles providing an effective normalization. For example, Lockwood et al. (2002) show that MSE for percentiles rapidly converges to a function that does not depend on K.

3.2 Loss for (below γ)/(above γ) classification

The USRDS uses percentiles to identify the best and the worst performers. Therefore, percentiles based on a loss function that specifically targets this goal should be evaluated. To this end, let $0 < \gamma < 1$ and consider classifying units as having a true percentile below or above 100γ and

also imposes a distance penalty. Assume that γK is an integer, so $\gamma(K+1)$ is not an integer and it is not necessary to make the distinction between $>$ and $\geq$ ($<$ and $\leq$). Among the loss functions considered by Lin et al. (2004) is:

$$\tilde{L}(\gamma) \;=\; K^{-1}\sum_k (\gamma - P_k^{est})^2\{I_{\{P_k>\gamma,\, P_k^{est}<\gamma\}} + I_{\{P_k<\gamma,\, P_k^{est}>\gamma\}}\}.$$

To minimize it, let $p_{k\ell} = pr(R_k = \ell \mid \mathbf{Y})$ and

$$\pi_k(\gamma) = pr(R_k \geq \gamma(K+1) \mid \mathbf{Y}) = \sum_{\ell=[\gamma K]+1}^{K} p_{k\ell}. \tag{3}$$

$\tilde{L}(\gamma)$ is minimized by:

$$\tilde{R}_k(\gamma) = \mathsf{rank}(\pi_k(\gamma));\; \tilde{P}_k(\gamma) = \tilde{R}_k(\gamma)/(K+1) \tag{4}$$

Dominici et al. (1999) use this approach with $\gamma = K/(K+1)$, ordering on the probability of the unit having the largest attribute.

3.3 **Ranking unit-specific summaries**

Ranks produced by ordering unit-specific summaries are more easily interpreted and computed. Candidate summaries include ρ_k^{mle}, ρ_k^{pm}, $\frac{\theta_k^{mle}}{sd(\theta_k^{mle})}$ and $pr(\rho_k > t \mid \mathbf{Y})$. The last computes an "exceedance probability"(Normand et al. (1997)). An interesting comparison with $\tilde{P}_k(\gamma)$ results by computing percentiles by ordering exceedance probabilities with threshold $t = \bar{G}_K^{-1}(\gamma)$ where $\bar{G}_K(t \mid \mathbf{Y}) = E_G[G_K(t;\boldsymbol{\rho}) \mid \mathbf{Y}] = \frac{1}{K}\sum_k pr(\rho_k \leq t \mid \mathbf{Y})$ (see Shen and Louis (1998)). We denote these by $P_k^*(\gamma)$ and they are monotone transform invariant.

4. Measures of uncertainty

Univariate or multivariate uncertainty summaries are available; we report on a selected set of univariate measures.

4.1 **Standard errors**

As for all statistical procedures, estimated ranks/percentiles must be accompanied by uncertainty measures. Using MCMC, any the standard errors of ranks produced by any ranking method can be computed. We report this computation for the SEL-minimizing $\bar{P}_k$ via their 95% CI. In addition, we compute the overall, posterior MSE for these estimates. If the data are completely uninformative, $MSE = 833, \sqrt{MSE} = 28.9$ and performance should be compared to these baselines.

4.2 **Operating characteristic for (below γ)/(above γ) classification**

The P_k^{est} from any ranking method can be used to classify units (below γ)/(above γ) . To compute an operating characteristic, define:

$$\begin{aligned} ABR(\gamma) &= pr(\text{a unit truly above } \gamma \text{ is classified below } \gamma \mid \mathbf{Y}) \\ BAR(\gamma) &= pr(\text{a unit truly below } \gamma \text{ is classified above } \gamma \mid \mathbf{Y}) \\ OC(\gamma) &= ABR(\gamma) + BAR(\gamma) = BAR(\gamma)/\gamma. \end{aligned} \tag{5}$$

If the goal is to identify units with the largest percentiles, then $BAR(\gamma)$ is similar to the False Discovery rate, Efron and Tibshirani (2002) Benjamini and Hochberg (1995), Storey (2002), Storey (2003). $ABR(\gamma)$ is similar to the False Non-Discovery Rate. The formula (5) for $OC(\gamma)$ results from the identity, $\gamma ABR(\gamma) = (1-\gamma)BAR(\gamma)$. When the data are completely uninformative ($gmv = \infty$ in the Gaussian case), $BAR(\gamma)/\gamma \doteq 1$ and so $OC(\gamma)$ produces a standardized comparison across cutpoints. Minimizing it produces the "most informative cutpoint."

Simulations can estimate the *a priori* $OC(\gamma)$; a Bayesian model structures computing the *a posteriori* values and using them as statistical summaries. To compute *a posteriori* values, let $\pi_k(\gamma) = pr(P_k > \gamma \mid \mathbf{Y})$ and, for a set of estimated percentiles P^{est}, define $\{k_j = k_j^{P^{est}} : P_k^{est} = j/(K+1)\}$. Then, suppressing dependency on $\mathbf{Y}$,

$$OC_{P^{est}}(\gamma) = BAR_{P^{est}}(\gamma)/\gamma \;=\; \frac{\sum_{j=[\gamma K]+1}^{K}[1-\pi_{k_j}(\gamma)]}{\gamma\{K-[\gamma K]\}}.$$

4.3 Performance curve for (below γ)/(above γ) classification

Plotting of $\pi_k(\gamma)$ versus P_k^{est} (see equation 3 and figure 2) gives the details on classification performance. (Similar plots can be constructed with the Y-axis being the exceedance probability.) Note that $OC(\gamma)$ is the area between $\pi_{k_j}(\gamma)$ and 1 for $j \geq [\gamma K] + 1$ plus the area below $\pi_{k_j}(\gamma)$ for $j \leq [\gamma K]$. Therefore, $OC_{P^{est}}(\gamma)$ is minimized by $P_k^{est} = \tilde{P}_k(\gamma)$.

4.4 Longitudinal variation

To measure variation in the ranks/percentile estimates within a dialysis center over the four years, we compute Longitudinal Variation, $LV(P^{est}) = 1000\frac{1}{3K}\sum_{k=1}^{K}\sum_{t=0}^{3}(P_{kt}^{est} - P_{k\cdot}^{est})^2$ where P_{kt}^{est} is the estimated percentile for unit k in year t and $P_{k\cdot}^{est}$ is the mean over the four years.

4.5 Subset dependency

Mathematical analyses show that ranks computed using the posterior distribution of the ranks are not subset invariant in that re-ranking the ranks for a subset of providers will not be the same as ranking only those providers. However, *if the prior distribution is known*, ranks based on unit-specific summaries such as the MLEs, PMs, exceedance probabilities or single-unit hypothesis tests are subset invariant. Of course, in a empirical Bayes or fully Bayesian analysis with an unknown prior (thus, including a hyper-prior), no method is subset invariant because the data are also used to estimate the prior or to update the hyper-prior. We investigate subset dependence by including/removing providers with small m_{kt} (high variance MLEs).

5. Results

5.1 Simulated Performance

We conducted simulation studies comparing ranking/percentiling methods for Poisson sampling distribution similar to those reported in Lin et al. (2004) for the Gaussian sampling distribution. Conclusions were similar with $\hat{P}_k$ performing well over a broad class of loss functions, with MLE-based ranks performing poorly, posterior mean-based ranks performing reasonably well but by no means optimal (see Louis and Shen (1999) and Gelman and Price (1999)). Performance of all

methods improved with increasing m_{kt} (reduced sampling variance), but performance being quite poor unless information in the sampling distribution is very high relative to that in the prior.

5.2 Subset dependency and the effect of unstable SMR estimates

We studied the effect including or excluding units with high-variance MLE estimates (small m_{kt}) by running both single-year and multiple-year analysis with and without the 68 providers with expected deaths < 0.1 in 1998. Comparisons based on $\hat{P}_k$ show that there is almost no change in percentiles for providers ranked either high or low, but there is noticeable re-ordering in the middle range. This is not surprising in that the ranks for high-variance providers are shrunken considerably towards the mid-rank $(K+1)/2$ and are not ranked at the extremes. The high variance providers "mix up" with the ranks from more stably estimated, central region providers, but are not contenders for extreme ranks/percentiles. Also, performance measures (MSE, $OC(\gamma)$) were very similar for the two datasets.

5.3 Comparisons using the 1998 data

We computed and compared estimates for 1998 using model 1 with $\phi \equiv 0$. Figure 1 displays relations similar to those in Conlon and Louis (1999). We display estimates for the 40 providers at the $1/3174, 82/3174, 163/3174, \ldots, 3173/3174$ percentiles as determined by $\hat{P}_k$. For each display, the Y-axis is $100\bar{P}_k$ with its 95% CI. By rows, the X-axis is $\hat{P}$, percentiles based on $E(\rho_k \mid \mathbf{Y})$, percentiles based on the MLEs of ρ and percentiles based on testing $\rho_k = 1$. To deal with small $Y_{kt} = 0$, for the hypothesis test statistic we use $\log(\frac{y_k}{m_k} + 0.25)\sqrt{m_k}$.

[Figure 1 about here.]

Note that in the upper left display the $\bar{P}_k$ do not fill out the (0, 100) percentile range; they are shrunken toward 50 by an amount that reflects estimation uncertainty. Also, the CIs are very wide, indicating considerable uncertainty in estimating percentiles. The plotted points are monotone because the X-axis is the ranked Y-axis values. Plotted points in the upper right display

are almost monotone; PM-based percentiles perform well. The lower left and lower right panel show considerable departure from monotonicity, indicating that MLE-based ranks and hypothesis test-based ranks are very far from optimal. Note also that the pattern of departures is quite different in the two panels, showing that these methods produce quite different ranks. Similar comparisons for more informative data (e.g., SMRs from the pooled 1998-2001 data) would be qualitatively similar, but the departures from monotonicity would be less extreme. See Lin et al. (2004) for additional comparisons using gene expression data.

5.4 Single year and multi-year analyses

Using model (1) we estimated single-year based and AR(1) model based percentiles. Table 1 reports that the ξ are near 0, as should be the case since we have used internal standardization, so the typical $\log(SMR) = 0$. The within year, between provider variation in $100\log(SMR)$ is essentially constant at approximately $100\tau = 24$, producing a 95% interval for true SMRs of (0.79, 1.27). Additional covariate adjustment could reduce this unexplained variation. The AR(1) model (with the posterior distribution for ϕ concentrated around 0.90) reduces $OC(0.8)$ by about 20% from about 61 to about 48. Classification performance using the $\hat{P}_k$ is very close to that for the optimal $\tilde{P}_k(0.8)$.

Longitudinal variation in ranks/percentiles (Longitudinal Variation, LV) is dramatically reduced for the AR(1) model going from 62 for the year-by-year analysis to 4 for the multi-year. As a basis for comparison, if $\phi \to 1$, $LV(\hat{P}) \to 0$ and if the data provide no information on the SMRs (the $\tau \to \infty$), then $LV(\hat{P}) = 83$.

[Table 1 about here.]

In Table 1, $100OC(0.8)$ is 62 and 49 for the single-year and AR(1) models. Figure 2 displays the details behind this superior classification performance. In the upper range of $\tilde{P}_k(\gamma)$, the curve for the AR(1) model lies above that for the single year, in the lower range it lies below. For the

AR(1) model to dominate the single year at all values of $\tilde{P}_k(\gamma)$, the curves would need to cross at $\tilde{P}_k(\gamma) = 0.8$, but the curves cross at about 0.7. We conjecture that if $m_{kt} \equiv m$, then the crossing would be at 0.8, but this remains to be investigated.

[Figure 2 about here.]

5.5 **Parametric and non-parametric priors**

We compare the parametric and NPML prior here based on data of 1998, i.e., $t = 0$ in single-year model 2. Figure 3 displays Gaussian, posterior expected and smoothed NPML estimated priors for $\theta = \log(\rho)$ using the 1998 data. The Gaussian is produced by plugging in the posterior median for (μ_0, τ_0). The posterior expected is a mixture of Gaussians using the posterior distribution of (μ_0, τ_0). The NPML is discrete and was smoothed using the "density" function in R with adjustment parameter = 10. The posterior distribution of (μ_0, τ_0) has close to 0 variance, so the two parametric curves superimpose. Note that the NPML has at least two modes with a considerable mass at approximately $\theta = 0.5; \rho = 1.65$. However, this departure from the Gaussian distribution has little effect on classification performance. Using 1998 data, for the NPML $100 \times OC(0.8) \approx 67$ while for the Gaussian prior the value is 62 (see Table 1). For performance evaluations of the NPML, see Paddock et al. (2004).

[Figure 3 about here.]

5.6 **Ranks based on Exceedance Probabilities**

Using the Gaussian prior for θ and the 1998 data, for $\gamma = 0.8$ the threshold $(\bar{G}_K^{-1}(\gamma))$ is $\theta = 0.169; \rho = 1.184$, indicating that the histogram of the unit-specific parameters is quite concentrated (as can be seen in Figure 3). The $P^*(0.8)$ are nearly identical to the $\tilde{P}(0.8)$ and the $(P_k^*, pr(\rho_k > 1.18))$ plot is virtually identical to the $\phi = 0$ curve in Figure 2. Additional study of these relations is needed.

6. Conclusion and discussion

A structured approach guided by a hierarchical model and a loss function is needed to produce ranks or percentiles that perform well. However, even optimal approaches can perform poorly and informative numerical and graphical performance assessments must accompany all estimates. Our assessments support those in Lin et al. (2004) regarding the generally good performance of $\hat{P}_k$, but also show that if a percentile cut-point γ can be identified, $\tilde{P}_k(\gamma)$ should be used. Our ensemble of performance measures (MSE, LV, OC) and graphical displays are but a subset of possible summaries and additional development is needed.

Ranks and percentiles computed through the posterior distribution of the ranks are prima facie relative comparisons. It is possible that all providers are doing well or that all are doing poorly and ranks won't pick this up. In situations where normative values are available (e.g., death rates), ranks that have a normative interpretation are attractive. (Of course, the SMR itself is a relative measure and so ranks produced from it are twice removed from a normative measure.) Ranking exceedance probabilities provides a monotone transform invariant procedure that provides a normative link. And, using as threshold the SMR value that is the γ^{th} percentile of the estimated cdf of SMR values (the P^*) produces ranks that are essentially identical to the $\tilde{P}_k(\gamma)$, thus connecting the latter to a normative measure.

Robustness of efficiency and validity are important attributes of any statistical procedure and basing assessments on the NPML or a more Bayesian alternative (see Paddock et al. (2004)) merits additional study and increased application.

Our approaches are based on loss functions that focus on a narrow aspect of performance assessment and broadening their purview will increase relevance. For example, in the USRDS application building in financial or other consequences of classification errors can help select γ and calibrate acceptable values of OC.

Finally, we need to educate stakeholders on the uses and abuses of ranks/percentiles; on their proper role in science and policy; on the absolute necessity of accompanying estimated ranks with uncertainty assessments and ensuring that these uncertainties influence decisions.

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, Series B, Methodological* **57**, 289–300.

Carlin, B. and Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall/CRC Press, Boca Raton, FL, 2^{nd} edition.

Christiansen, C. and Morris, C. (1997). Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* **127**, 764–768.

Conlon, E. and Louis, T. (1999). Addressing multiple goals in evaluating region-specific risk using Bayesian methods. In Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F. and Bertollini, R., editors, *Disease Mapping and Risk Assessment for Public Health*, chapter 3, pages 31–47. Wiley.

Dominici, F., Parmigiani, G., Wolpert, R. L. and Hasselblad, V. (1999). Meta-analysis of migraine headache treatments: Combining information from heterogeneous designs. *Journal of the American Statistical Association* **94**, 16–28.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–139.

Efron, B. and Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.

ESRD (2000). 1999 annual report: ESRD clinical performance measures project. Technical report, Health Care Financing Administration.

Gelman, A. and Price, P. (1999). All maps of parameter estimates are misleading. *Statistics in Medicine* **18**, 3221–3234.

Goldstein, H. and Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society Series A* **159**, 385–443.

Kendziorski, C., Irizarry, R. A., Chen, K., Haag, J. and Gould, M. (2004). To pool or not to pool: A question of microarray experimental design. *PNAS* Under review. `http://www.bepress.com/jhubiostat/paper46`.

Lacson, E., Teng, M., Lazarus, J., Lew, N., Lowrie, E. and Owen, W. (2001). Limitations of the facility-specific standardized mortality ratio for profiling health care quality in dialysis. *American Journal of Kidney Diseases* **37**, 267–275.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **78**, 805–811.

Landrum, M., Bronskill, S. and Normand, S.-L. (2000). Analytic methods for constructing cross-sectional profiles of health care providers. *Health Services and Outcomes Research Methodology* **1**, 23–48.

Lin, R., Louis, T., Irizarri, R. and Parmigiani, G. (2004). Gene identifcation using optimal ranking methods. Technical report, Department of Biostatistics, Johns Hopkins SPH.

Lin, R., Louis, T. A., Paddock, S. M. and Ridgeway, G. (2004). Loss function based ranking in two-stage, hierarchical models. Technical report, Johns Hopkins University, Dept. of Biostatistics Working Papers.

Liu, J., Louis, T., Pan, W., Ma, J. and Collins, A. (2004). Methods for estimating and interpreting provider-specific, standardized mortality ratios. *Health Services and Outcomes Research Methodology* **4**, 135–149.

Lockwood, J., Louis, T. and McCaffrey, D. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics* **27**, 255–270.

Louis, T. and Shen, W. (1999). Innovations in Bayes and empirical Bayes methods: Estimating parameters, populations and ranks. *Statistics in Medicine* **18**, 2493–2505.

McClellan, M. and Staiger, D. (1999). The quality of health care providers. Technical Report 7327, National Bureau of Economic Research, Working Paper.

Newton, M., Kendziorski, C., Richmond, C., Blatterner, F. and Tsui, K. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

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.

Paddock, S., Ridgeway, G., Lin, R. and Louis, T. (2004). Flexible prior distributions and triple goal estimates in two-stage, hierarchical linear models. Technical report, Department of Biostatistics, Johns Hopkins SPH.

Shen, W. and Louis, T. (1998). Triple-goal estimates in two-stage, hierarchical models. *Journal of the Royal Statistical Society, Series B* **60**, 455–471.

Shen, W. and Louis, T. (2000). Triple-goal estimates for disease mapping. *Statistics in Medicine* **19**, 2295–2308.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B, Methodological* **64**, 479–498.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value1. *The Annals of Statistics* **31**, 2013–2035.

USRDS (2003). 2003 annual data report: Atlas of end-stage renal disease in the United States. Technical report, Health Care Financing Administration.

Zaslavsky, A. M. (2001). Statistical issues in reporting quality data: Small samples and casemix variation. *International Journal for Quality in Health Care* **13**, 481–488.

Appendix: The NPML

Assume $\rho_k \sim G, k = 1, \ldots, K$ with G discrete having at most J mass points $u_1, \ldots, u_J$ with probabilities $p_1, \ldots, p_J$. To estimate the us and ps, start with $u_1^{(0)}, \ldots, u_J^{(0)}$ and $p_1^{(0)}, \ldots, p_J^{(0)}$ and use the EM algorithm, for the recursion,

$$\begin{aligned} w_{kj}^{(v+1)} &= \mathsf{pr}(\rho_k = u_j^{(v)} | \mathsf{data}) \\ w_{kj}^{(v+1)} &= \frac{(m_k u_j^{(v)})^{y_k} e^{-m_k u_j^{(v)}} p_j^{(v)}}{\sum_l (m_k u_l^{(v)})^{y_k} e^{-m_k u_l^{(v)}} p_j^{(v)}} \\ p_j^{(v+1)} &= \frac{w_{+j}^{(v+1)}}{w_{++}^{(v+1)}} \\ u_j^{(v+1)} &= \frac{\sum_k w_{kj}^{(v+1)} y_k}{\sum_k w_{kj}^{(v+1)} m_k}. \end{aligned} \tag{6}$$

This recursion converges to a fixed point $\widehat{G}$ and, if unique, to the NPML. The recursion is stopped when the maximum relative change in each step for both the $u_j^{(v)}$ and the $p_j^{(v)}$, $j = 1, 2, \cdots, K$ is smaller than 0.001. At convergence, $\widehat{G}$ is both prior and the Shen and Louis (1998) histogram estimate $\widehat{G}_K$.

Care is needed in programming the recursion. The w-recursion is:

$$w_{kj}^{(v+1)} = \frac{(m_k u_j^{(v)})^{y_k} e^{-m_k u_j^{(v)}} p_j^{(v)}}{\sum_l (m_k u_l^{(v)})^{y_k} e^{-m_k u_l^{(v)}} p_j^{(v)}}.$$

Since $e^{-m_k u_j^{(v)}}$ can be extremely small ($m_k u_j^{(v)}$ can be extremely large), to stabilize the computations we define,

$$\overline{\rho}^{(v)} = \sum_j p_j(v) u_j^{(v)},$$

and write

$$\left(m_k u_j^{(v)}\right)^{y_k} = e^{y_k \log\ m_k u_j^{(v)}} \quad .$$

The w-recursion becomes:

$$\begin{aligned} w_{kj}^{(v+1)} &= \frac{(u_j^{(v)}/\overline{\rho}^{(v)})^{y_k} e^{-m_k(u_j^{(v)}-\overline{\rho}^{(v)})} p_j^{(v)}}{\sum_{l=1}^{J}(u_l^{(v)}/\overline{\rho}^{(v)})^{y_k} e^{-m_k(u_l^{(v)}-\overline{\rho}^{(v)})} p_l^{(v)}} \\ &= \frac{p_j^{(v)} e^{(y_k \log(u_j^{(v)}/\overline{\rho}^{(v)}) - m_k(u_j^{(v)}-\overline{\rho}^{(v)}))}}{\sum_{l=1}^{J} p_l^{(v)} e^{(y_k \log(u_l^{(v)}/\overline{\rho}^{(v)}) - m_k(u_l^{(v)}-\overline{\rho}^{(v)}))}} \end{aligned}$$

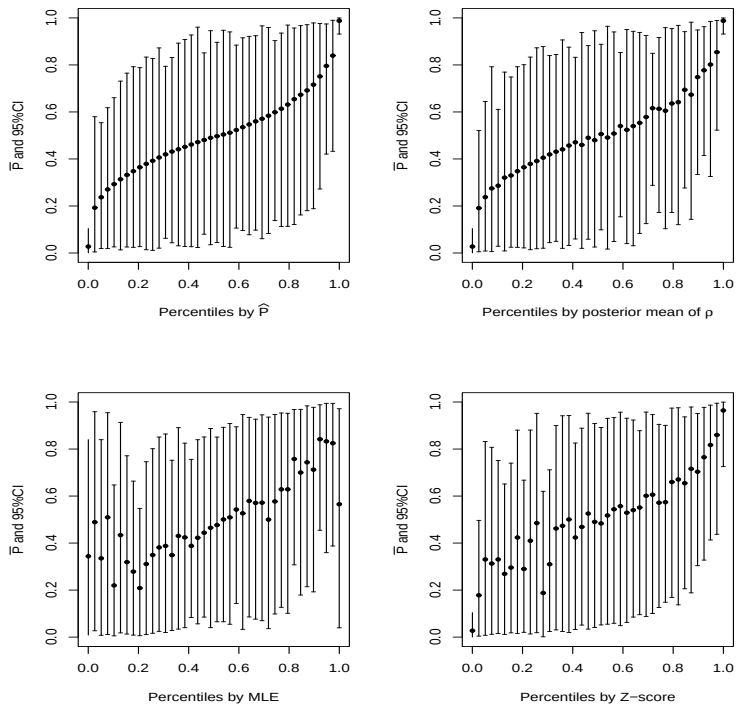

Figure 1. SEL-based percentiles for 1998. For each display, the Y-axis is $100\bar{P}_k$ with its 95% CI. By rows, the X-axis is $\hat{P}$, percentiles based on $E(\rho_k \mid \mathbf{Y})$, percentiles based on the MLEs of ρ and percentiles based on testing $\rho_k = 1$.

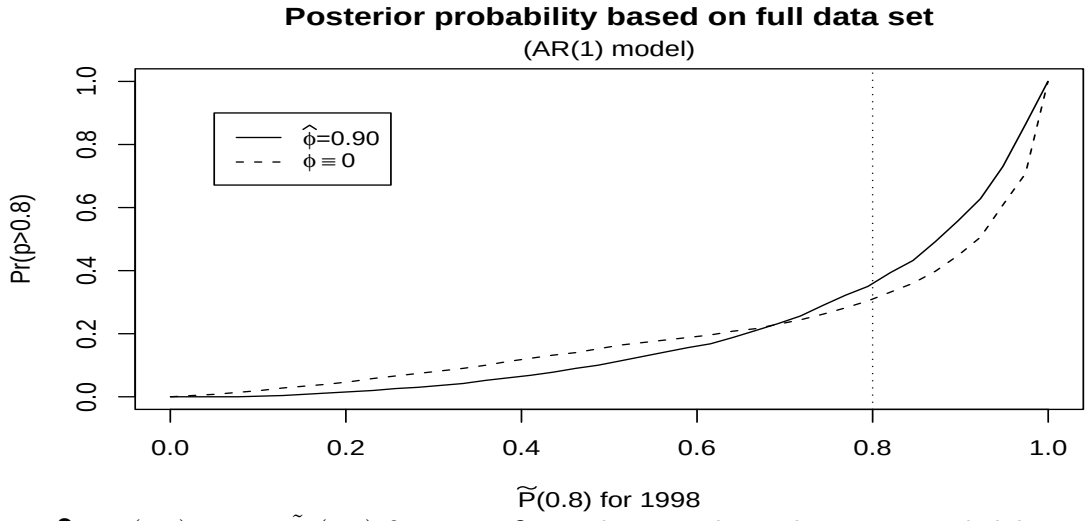

Figure 2. $\pi_k(0.8)$ versus $\tilde{P}_k(0.8)$ for 1998. Optimal percentiles and posterior probabilities computed by the single year model ($\phi \equiv 0$) and the AR(1) model ($\hat{\phi} = 0.90$).

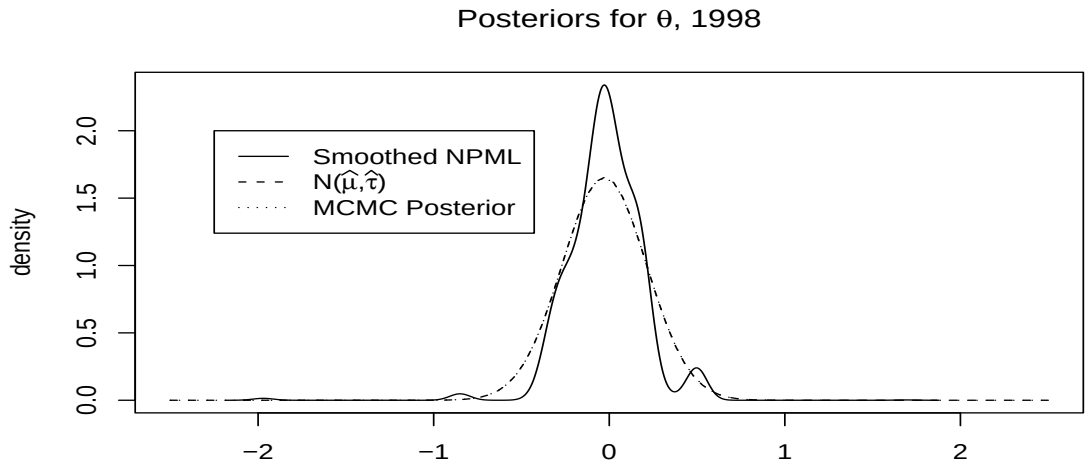

Figure 3. Estimated priors for $\theta = \log(\rho)$ using the 1998 data. The dashed curve is Gaussian with posterior medians for (μ, τ); the dotted curve is a mixture of Gaussians using the posterior distribution of (μ, τ); the solid curve is a smoothed NPML using the "density" function in R with adjustment parameter $= 10$.

	Single Year: ($\phi \equiv 0$)				Multi-Year: ($\boldsymbol{100\phi \sim {}_{88}90_{92}}$)			
Parameter	1998	1999	2000	2001	1998	1999	2000	2001
100ξ	-2.8	-1.3	-2.3	-0.7	-3.1	-0.8	-1.7	-0.3
100τ	24.1	23.5	23.1	22.2	25.8	25.0	24.9	24.1
$100 \times OC_{\tilde{P}(0.8)}(0.8)$	62	61	60	62	49	47	46	50
LV($\hat{P}_k$)	62				4			

Table 1

Data analysis results for $\hat{P}_k$ and $\tilde{P}(0.8)$. In the multi-year section, $100OC(0.8)$ is for the indicated year as estimated from the multi-year model and ${}_{88}90_{92}$ is the posterior median and 95% credible interval for 100ϕ.