Published in final edited form as:
Health Serv Outcomes Res Methodol. 2009 March 1; 9(1): 22–38. doi:10.1007/s10742-008-0040-0.

Ranking USRDS provider specific SMRs from 1998-2001

Rongheng Lin
Department of Public Health, University of Massachusetts Amherst, Rm 411 Arnold House, 715 N. Pleasant Rd., Amherst, MA 01003, USA

Thomas A. Louis
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA e-mail: tlouis@jhsph.edu

Susan M. Paddock
RAND Corporation, Santa Monica, CA 90407, USA e-mail: paddock@rand.org

Greg Ridgeway
e-mail: gregr@rand.org

Abstract

Provider profiling (ranking/percentiling) is prevalent in health services research. Bayesian models coupled with optimizing a loss function provide an effective framework for computing non-standard inferences such as ranks. Inferences depend on the posterior distribution and should be guided by inferential goals. However, even optimal methods might not lead to definitive results and ranks should be accompanied by valid uncertainty assessments. We outline the Bayesian approach and use estimated Standardized Mortality Ratios (SMRs) in 1998-2001 from the United States Renal Data System (USRDS) as a platform to identify issues and demonstrate approaches. Our analyses extend Liu et al. (2004) by computing estimates developed by Lin et al. (2006) that minimize errors in classifying providers above or below a percentile cut-point, by combining evidence over multiple years via a first-order, autoregressive model on log(SMR), and by use of a nonparametric prior. Results show that ranks/percentiles based on maximum likelihood estimates of the SMRs and those based on testing whether an SMR = 1 substantially under-perform the optimal estimates. Combining evidence over the four years using the autoregressive model reduces uncertainty, improving performance over percentiles based on only one year. Furthermore, percentiles based on posterior probabilities of exceeding a properly chosen SMR threshold are essentially identical to those produced by minimizing classification loss. Uncertainty measures effectively calibrate performance, showing that considerable uncertainty remains even when using optimal methods. Findings highlight the importance of using loss function guided percentiles and the necessity of accompanying estimates with uncertainty assessments.

Keywords

Provider profiling; Ranks/percentiles; Bayesian hierarchical model; Uncertainty assessment

1 Introduction

Research on and application of performance evaluation steadily increases with applications to evaluating health service providers (Christiansen and Morris 1997; Goldstein and Spiegelhalter

Correspondence to: Rongheng Lin.
e-mail: rlin@schoolph.umass.edu.

1996; Landrum et al. 2000; Liu et al. 2004; McClellan and Staiger 1999; Grigg et al. 2003; Zhang et al. 2006; Normand and Shahian 2007; Ohlssen et al. 2007), prioritizing environmental assessments in small areas (Conlon and Louis 1999; Louis and Shen 1999; Shen and Louis 2000) and ranking teachers and schools (Lockwood et al. 2002). Inferential goals of these studies include evaluating the population performance, such as the average performance of all health providers and comparing performance among providers. Performance evaluations include comparing unit-specific, substantive measures such as death rates, identifying the group of poorest or best performing units and overall ranking of the units, e.g., profiling or league tables (Goldstein and Spiegelhalter 1996).

The Standardized Mortality Ratio (SMR), the ratio of observed to expected deaths, is an important service quality indicator (Zaslavsky 2001). The United States Renal Data System (USRDS) produces annual estimated SMRs for several thousand dialysis centers and uses these as a quality screen (Lacson et al. 2001; ESRD 2000; USRDS 2005). Invalid estimation or inappropriate interpretation can have serious consequences for these dialysis centers and for their patients. We present an analysis of the information from the United States Renal Data System (USRDS) for 1998-2001 as a platform for demonstrating and comparing approaches to ranking health service providers. From the USRDS we obtained observed and expected deaths for the $K = 3173$ dialysis centers that contributed information for all four years. The approach used by USRDS to produce these values can be found in USRDS (2005).

Though estimating SMRs is a standard statistical operation (produce provider-specific expected deaths based on a statistical model, and then compute the “observed/expected” ratio), it is important and challenging to deal 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), and the need to adjust for potential biases induced when attributing deaths to providers and accounting for informative censoring.

The multi-level data structure and complicated inferential goals require the use of a hierarchical Bayesian model that accounts for nesting relations and specifies both population values and random effects. Correctly specified, the model properly accounts for the sample design, variance components and other uncertainties, producing valid and efficient estimates of population parameters, variance components and unit-specific random effects (provider-, clinician-, or region-specific latent attributes), all accompanied by valid uncertainty assessments. Importantly, the Bayesian approach provides the necessary structure for developing scientific and policy-relevant inferences based on the joint posterior distribution of all unknowns.

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 Bayesian approach structures non-standard inferences such as ranking (including identification of extremely poor and good performers) and estimating the histogram of unit-specific random effects. For example, as Liu et al. (2004) show, when estimation uncertainty varies over dialysis centers, ranks produced by Z-scores that test whether a provider's SMR = 1 tend to identify providers with relatively low variance as extreme because these tests have the highest power; ranks produced from the provider-specific maximum likelihood estimates (MLEs) are more likely to identify dialysis centers with relatively high variance as extreme. Effective ranks depend on striking an effective tradeoff between signal and noise.

Lin et al. (2006) present estimates that minimize errors in classifying providers above or below a percentile cut-point. Our analyses build on Liu et al. (2004) by extending the application of

Lin et al. (2006)'s estimates to combine evidence over multiple years via a first-order, autoregressive model on log(SMR), and by use of a nonparametric prior. For single-year analyses we compare the results from the log-normal prior to those based on the Non-Parametric, Maximum Likelihood (NPML) prior (Laird 1978).

In following, Sect. 2 presents our models; Sect. 3 outlines several ranking methods; Sect. 4 gives uncertainty measures; Sect. 5 presents results and Sect. 6 sums up and identifies additional research. Computing code for all routines is available at, http://people.umass.edu/rlin/jhuwebhost/usrds-ranking.htm.

2 Statistical models

We employ both single-year and longitudinal models for observed deaths and underlying parameters, with the former a sub-model of the latter. To this end, let (Y_{kt}, m_{kt}) be the observed and case-mix adjusted, expected deaths for provider k in year t, k = 1,... 3173, t = 0, 1, 2, 3 and ρ_{kt} be the SMR. The USRDS computes the expecteds under the assumption that all providers give the same quality of care for patients with identical covariates, see USRDS (2005) for details. We employ the conditional Poisson model,

$$\begin{aligned} Y_{kt}|m_{kt},\rho_{kt} &\sim \text{Poisson}(\rho_{kt}m_{kt}) \\ E(Y_{kt}|m_{kt},\rho_{kt}) &= m_{kt}\rho_{kt}. \end{aligned} \tag{1}$$

If the provider has “average performance”, ρ_{kt} = 1. For both single-year and multiple-year analyses we model $\theta_{kt} = \log(\rho_{kt})$.

2.1 Single-year analyses

For single-year analyses, we assume that for year t; $\theta_{kt} \overset{iid}{\sim} G_t$, k = 1,…, 3173: We use a year-specific, normal prior (see the note after Eq. 2) and for the single-year analyses also use the non-parametric maximum likelihood (NPML) prior. See and Carlin and Louis (2008) and Paddock et al. (2006) for additional details and Appendix C for the estimation algorithm.

2.2 The longitudinal, *AR*(1) model

To model longitudinal correlation among (ρ_{k0}, ρ_{k1}, ρ_{k2}, ρ_{k3}), let $\phi = \text{cor}(\theta_{k,t}, \theta_{k(t+1)})$, with -1 < ϕ < 1. Then, use a normal prior on the θ_{kt} and a normal prior on $Z(\phi) = 0.5 \log\{(1+\phi)/(1-\phi)\}$ in the hierarchical model,

$$\begin{gathered} \xi_t \overset{iid}{\sim} N(0,V), \quad \lambda_t = \tau_t^{-2} \overset{iid}{\sim} \text{Gamma}(\alpha,\mu/\alpha) \\ Z(\phi) \sim N\left(0,V_\phi\right) \\ [\theta_{10},\ldots,\theta_{K0}|\xi_0,\tau_0] \overset{iid}{\sim} N\left(\xi_0,\tau_0^2\right) \\ [\theta_{kt}|\theta_{k(t-1)},\ldots,\theta_{k0},\xi,\tau,\phi] \overset{ind}{\sim} N\left(\xi_t+\phi\tau_t\tau_{t-1}^{-1}\{\theta_{k(t-1)}-\xi_{t-1}\},\left\{1-\phi^2\right\}\tau_t^2\right) \\ [Y_{kt}|m_{kt},\rho_{kt}] \sim \text{Poisson}(m_{kt}\rho_{kt}), \rho_{kt} = \exp(\theta_{kt}). \end{gathered} \tag{2}$$

The notation “iid” means independently and identically distributed and “ind” means independently distributed. The relation is first-order Markov, because though conditioning is on all prior θs, only $\rho_{k(t-1)}$ appears on the right-hand side of Eq. 2.

Marginally, for year t, θ_{kt} iid $N(\xi_t, \tau_t^2)$ and setting ϕ = 0 produces four, single-year analyses, each using the Liu et al. (2004) model with no borrowing of information over time. For ϕ > 0,

we have a standard AR(1) model on the latent log(SMR)s and the posterior distribution combines evidence across dialysis centers within year and within dialysis center across years.

2.3 Posterior sampling implementation and hyper-prior parameters

We implement a Gibbs sampler for model (2) with WinBUGS via the R package *R2WinBUGS*, using the *coda* package to diagnose convergence (Spiegelhalter et al. 1999; Gelman et al. 2006; Plummer et al. 2006). We use $V = 10$, $\mu = 0.01$, $\alpha = 0.05$, values that stabilize the simulation while allowing sufficient adaptation to the data. With $V = 10$, the *a priori*, 95% probability interval for ξ_t is (-6.20, 6.20) [(0.002, 492.75) in the SMR scale]; the values for α and μ produce a distribution for τ^2 with center near 100, inducing large, *a priori* variation for the θ_{kt}. For the *AR*(1) model, reported results are based on the $V_\phi = 0.2$. This produces an *a priori* 95% probability interval for ϕ of (-0.70, 0.70). In a sensitivity analysis, we also tried $V_\phi = 2$, which produced the *a priori* interval (-0.99, 0.99) and yielded results virtually identical to those based on the $V_\phi = 0.2$ hyper-prior. In both cases, the data likelihood dominated the priors. This can also be seen in the shrinkage of τ^2 towards zero, as reported in the Sect. 5.4. There is no strong posterior correlation observed between ϕ and the τ^2s.

3 Loss function based ranking methods

Two general strategies for ranking are available. The preferred strategic approach first computes the joint posterior distribution of the ranks and then uses it to produce estimates and uncertainty assessments, generally guided by a loss function that is appropriate for analytic goals. This approach ensures that estimated ranks have desired properties such as not depending on a monotone transform of the target parameters. The other approach is based on ordering estimates of target parameters (MLEs or posterior means) or on ordering statistics testing the null hypothesis that $SMR_k \equiv 1$. If the posterior distributions of the target parameters are stochastically ordered, then for a broad class of loss functions (estimation criteria) optimally estimated ranks will not depend on the strategy. However, Lin et al. (2006) and others have shown that estimates not derived from the distribution of the ranks can perform very poorly and may not be invariant under monotone transformation of the target parameters. Producing the joint posterior distribution of the ranks is computationally intensive, but most estimates depend only on easily computable features.

We first define ranks and then specify candidate ranking methods. For clarity in defining ranks, we drop the index *t* and write $R_k(\rho) = \text{rank}(\rho_k) = \sum_{j=1}^{K} I_{\{\rho_k \geq \rho_j\}}$, with the smallest ρ_k having rank 1. Rank-based estimates are based on the joint posterior distribution of the $R_k(\rho)$ and are invariant under monotone transform of the ρ_k.

3.1 Squared-error loss

Shen and Louis (1998) and Lockwood et al. (2002) study ranks that minimize the posterior risk induced by squared error loss (SEL): $\mathrm{E}\left[K^{-1}\sum_k \left(R_k^{est} - R_k(\rho)\right)^2\right]$. It is minimized by the posterior expected ranks,

$$\bar{R}_k(\mathbf{Y}) = E_{\rho|\mathbf{Y}}\left[R_k(\rho)|\mathbf{Y}\right] = \sum_{j=1}^{K} \mathrm{pr}\left(\rho_k \geq \rho_j|\mathbf{Y}\right), \tag{3}$$

where pr(·) stands for probability. The optimal mean squared error (MSE) in estimating the ranks is equal to the average posterior variance of the ranks. Generally, the $\bar{R}_k$ are not integers; for optimal, distinct integer ranks, use $\widehat{R}_k(\mathbf{Y}) = \text{rank}\left(\bar{R}_k(\mathbf{Y})\right)$.

In the notation that follows, generally we drop dependency on ρ (equivalently, on θ) and omit conditioning on **Y**. For example, R_k stands for $R_k(\theta)$ and $\widehat{R}_k$ stands for $\widehat{R}_k(\theta|\mathbf{Y})$. 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 of ranking estimator and posterior distributions of parameters that does not depend on K.

3.2 Optimizing (above γ)/(below γ) classification errors

The USRDS uses percentiles to identify the best and the worst performers. Let γ be the fraction of top performers among the total that we want to identify, $0 < \gamma < 1$. A loss function designed to address this inferential goal was proposed by Lin et al. (2006). The loss function (Eq. 4) penalizes for misclassification and also imposes a distance penalty between estimated percentiles and the cutoff γ.

$$\tilde{L}(\gamma) = K^{-1}\sum_k \left(\gamma - P_k^{est}\right)^2 \left\{ I_{\left\{P_k>\gamma,\quad P_k^{est}<\gamma\right\}} + I_{\left\{P_k<\gamma,\quad P_k^{est}>\gamma\right\}} \right\} \tag{4}$$

For ease of presentation, we have assumed that γK is an integer and so $\gamma(K + 1)$ is not. It is not necessary to make the distinction between > and ≥. To minimize the posterior risk induced by (4), let $p_{k\ell} = \text{pr}(R_k = \ell|\mathbf{Y})$ and $\pi_k(\gamma) = \text{pr}(R_k > \gamma(K+1)|\mathbf{Y}) = \sum_{\ell=[\gamma K]+1}^{K} p_{k\ell}$.

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

$$\begin{aligned} \tilde{R}_k(\gamma) &= \text{rank}(\pi_k(\gamma)) \\ \tilde{P}_k(\gamma) &= \tilde{R}_k(\gamma)/(K+1) \end{aligned} \tag{5}$$

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

3.3 Equivalence of the $\tilde{P}_k(\gamma)$ and ordering posterior exceedance probabilities

Given an SMR threshold t, the ranks/percentiles induced by ordering the posterior probabilities that an SMR exceeds the threshold, $\text{pr}(\rho_k > t|\mathbf{Y})$ allow us to make a connection between the $\tilde{P}_k(\gamma)$ and the substantive scale (in our application, SMR). Normand et al. (1997) rank providers based on these "exceedance probabilities" and Diggle et al. (2007) use them to identify the areas with elevated disease rates. Lin et al. (2006) shows that exceedance probability based percentiles are virtually identical to the $\tilde{P}_k(\gamma)$ by choosing the γth percentile of the average of posterior cumulative distribution function as the threshold t, i.e., $t = \bar{G}_K^{-1}(\gamma)$, where $\bar{G}_k(t) = \frac{1}{K}\sum_k \text{pr}(\rho_k < t|\mathbf{Y})$. We denote the percentiles based on $\text{pr}\left(\rho_k > \bar{G}_K^{-1}(\gamma)|\mathbf{Y}\right)$ as $P_k^*(\gamma)$. In addition to providing a connection to the SMR scale, the $P_k^*(\gamma)$ are far easier to compute than are the $\tilde{P}_k(\gamma)$. Note that the $P_k^*(\gamma)$ are invariant under the monotone transform of ρ_k.

4 Performance measures

As for all statistical procedures, estimated ranks/percentiles must be accompanied by uncertainty statement. A wide variety of univariate and multivariate performance measures are available and we propose three univariate measures of uncertainty.

4.1 Mean squared error

Using MCMC, the posterior mean squared error of percentiles produced by any method and 95% posterior intervals of dialysis center-specific percentiles can be computed. As a baseline, if the data are completely uninformative so that the percentiles $\widehat{P}_k$ (1/3174, 2/3174,…, 3173/3174) are randomly assigned to the 3173 dialysis centers, then $10000 \times MSE = 1666, 100 \times \sqrt{MSE} = 41$.

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

The vector of P_k^{est} from any ranking method can be used to classify units into (above γ)/(below γ) groups and the posterior classification performance (operating characteristic) can be computed. Following Lin et al. (2006), and suppressing dependence on P^{est}

$$OC(\gamma|\mathbf{Y}) = ABR(\gamma|\mathbf{Y}) + BAR(\gamma|\mathbf{y}) = BAR(\gamma|\mathbf{Y})/\gamma \tag{6}$$

where, $ABR(\gamma|\mathbf{Y})$ = pr(percentile) > γ|percentile estimated < γ, **Y**) = pr($P > \gamma | P^{est} < \gamma$, **Y**) $BAR(\gamma|\mathbf{Y})$ = pr(percentile < γ|percentile estimated > γ, **Y**) = pr($P < \gamma | P^{est} > \gamma$, **Y**).

The second equality in (6) results from the identity, $\gamma ABR(\gamma|\mathbf{Y}) = (1 - \gamma) BAR(\gamma|\mathbf{Y})$. If the goal is to identify units with the largest percentiles, then $BAR(\gamma|\mathbf{Y})$ is similar to the False Discovery Rate (Benjamini and Hochberg 1995; Efron and Tibshirani 2002; Storey 2002; Storey 2003). $ABR(\gamma|\mathbf{Y})$ is similar to the False non-Discovery Rate. When the data are completely uninformative, $BAR(\gamma|\mathbf{Y})/\gamma \doteq 1$ and so $OC(\gamma|\mathbf{Y})$ produces a standardized comparison across γ values. Minimizing it produces the most informative cut point for a given P^{est}.

For any percentiling method, $OC(\gamma|\mathbf{Y})$ provides a data analytic performance evaluation. The direct computation of it sums $\pi_k(\gamma|\mathbf{Y})$ = pr($P_k > (\gamma|\mathbf{Y})$ over a P^{est} produced set of indices.

$$OC_{P^{est}}(\gamma|\mathbf{Y}) = BAR_{P^{est}}(\gamma|\mathbf{Y})/\gamma = \frac{\sum_{j=1}^{K} [1 - \pi_k(\gamma|\mathbf{Y})] I_{\{P_k^{est} > \gamma\}}}{\gamma \{K - [\gamma K]\}}.$$

Plotting the $\pi_k(\gamma|\mathbf{Y})$ versus the P_k^{est} (see Fig. 1) displays percentile-specific, classification performance. For ideal fully informative data, the exceedance probability should be 1 for those classified as above γ and 0 for those classified as below γ. $OC(\gamma)$ is the area between $\pi_{k_j}(\gamma)$ curve and 1 for $j \geq [\gamma K] + 1$ plus the area below $\pi_{k_j}(\gamma)$ curve for $j \leq [\gamma K]$. Using $\tilde{P}_k(\gamma)$ for the X-axis produces a monotone plot and $OC_{\tilde{P}_k(\gamma)}(\gamma)$ is the minimum attainable. This plot is similar to that proposed by Pepe et al. (2008)

Computing the $\pi_k(\gamma|\mathbf{Y})$ is numerically challenging. However, the virtual equivalence between $\tilde{P}_k(\gamma)$ and $P_k^*(\gamma)$ justifies replacing these posterior probabilities by the easily computed pr($\rho_k > t|\mathbf{Y}$) with $t = \bar{G}_K^{-1}(\gamma)$.

4.3 Longitudinal variation

For most of dialysis centers, we expect that percentile estimates of different years are similar to each other. To measure variation in the ranks/percentile estimates within dialysis centers over the four years, we compute Longitudinal Variation:

$$LV_{P^{est}} = 1000 \times \frac{1}{3K} \sum_{k=1}^{K} \sum_{t=0}^{3} \left(P_{kt}^{est} - P_{k\cdot}^{est} \right)^2,$$

where P_{kt}^{est} is the estimated percentile for dialysis center k in year t and $P_{k\cdot}^{est}$ is the mean over the four years. A smaller LV value indicates better consistency in percentiles estimates of different years.

4.4 Subset dependency

Unlike estimating individual parameters (where there is individual shrinkage), ranks are highly correlated and so changing the posterior distribution of some target parameters or removing or adding units rearrange the order of individual parameters in a complicated manner. Ranks computed using the posterior distribution of the ranks are thus not subset invariant in that re-ranking the ranks for a subset of providers will not be the same as ranking only those providers. Section Appendix A gives a numeric illustrative example. However, *if the prior distribution is known*, ranks based on provider-specific summaries such as the MLEs, PMs, exceedance probabilities or single-provider hypothesis tests are subset invariant. Of course, in an empirical Bayes or fully Bayesian analysis with an unknown prior (thus, including a known 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). These providers are generally small dialysis centers with very few patients. Ranking procedures excluding these centers imply that the centers are first categorized according to their sizes and rankings are then generated in different categories separately. We pursue our comparison under model (2).

5 Results

5.1 Simulated performance

We conducted simulation studies comparing ranking/percentiling methods for the Poisson sampling distribution similar to those reported in Lin et al. (2006) for the Gaussian sampling distribution. Conclusions were similar with $\widehat{P}_k$ performing well over a broad class of loss functions, with MLE-based ranks performing poorly and ranks of posterior mean performing reasonably well but by no means optimally (see Louis and Shen 1999; Gelman and Price 1999). Performances of all methods improved with increasing m_{kt} (reduced sampling variance), but generally the ranking results are quite indefinitive unless information in the sampling distribution (e.g., provided by the data) 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 providers with small m_{kt} (high-variance MLE estimates) 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 $\widehat{P}_k$ in a graph similar to Fig. 2 shows that, there is almost no change in percentiles for providers ranked either high or low, but noticeable re-ordering happens in the middle range. This is not surprising in that the ranks for high-variance providers are shrunken considerably towards the midrank $(K + 1)/2$ 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, there are more providers in a given interval length in the middle of the distribution of parameters than in the tails. The ranks of these large providers in the middle range will be more sensitive to the change of the joint posterior distribution caused by including small providers. Performance measures MSE and $OC(\gamma)$ were very similar for the two datasets.

5.3 Comparisons using the 1998 data

We computed ranks (formula (7)) based on the MLE and hypothesis testing statistics Z-scores (testing the hypothesis H_0: ρ = 1 for 1998); we computed the Bayesian estimates $\widehat{P}_k, \tilde{P}_k(\gamma)$ and percentiles based on the posterior means (ρ_k^{pm}) using model (2) with $\phi \equiv 0$.

$$\rho_k^{mle} = \frac{Y_k}{m_k}, \quad Z_k = \sqrt{m_k}\log\left(\frac{y_k}{m_k}\right), \quad \rho_k^{pm} = E(\rho_k|\mathbf{Y}). \tag{7}$$

Globally, if we regard a dialysis center with ρ_k^{mle} greater than 1.5 as "flagged," then 379 (12%) of dialysis centers will be identified; if we regard a dialysis center with Z-score greater than 1.645 as "flagged," then 647 (20.4%) of dialysis centers are identified.

To compare methods, we select the 634 (20%) worst dialysis centers by ranking and selecting the largest MLEs and Z-scores and compare to those identified by $\tilde{P}(0.8)$. The 80th percentiles of the ρ^{mle} and Z-score are 1.44 and 1.67, respectively, whereas the 80th percentile of the ρ^{pm} is 1.10 (these PMs are closer to 1 than their respective MLEs).

We calculate the kappa statistics between the (above γ)/(below γ) classifications based on $\tilde{P}_k(\gamma)$ and other estimators. The classifications based on $\widehat{P}_k$ and ρ^{pm} have high agreement with those based on $\tilde{P}_k(\gamma)$ with respective kappa statistics 0.90 and 0.94. The kappa statistics between the MLE and $\tilde{P}_k(\gamma)$ is 0.78, between the Z-score and $\tilde{P}_k(\gamma)$ is 0.83, and between MLE and Z-score is the lowest 0.68.

Figure 3 compares different methods based on their posterior probability of correct classification $\text{pr}(P_k > 0.8|\mathbf{Y})$ The curve for $\tilde{P}_k(\gamma)$ is monotone and optimal because we construct $\tilde{P}_k(\gamma)$ by ranking these probabilities. The curves for percentiles based on ρ^{pm} and on $\widehat{P}_k$ are very close to that for $\tilde{P}_k(\gamma)$ (not plotted). The curves for MLE-based and Z-score-based percentiles are far from monotone.

The MLE SMRs for centers with relatively small expected deaths have relatively large variances. To study the impact of large and small variances on estimated percentiles, Fig. 3 identifies those for the 147 dialysis centers that treated fewer than 5 patients in 1998. These constitute 4.5% of all centers. Generally, MLE-based percentiles for these centers are at the extremes whereas Z-score based percentiles tend to be near 0.5. However, because the posterior distribution of ρ_k for the high variance centers is concentrated around 1, the $\tilde{P}_k(\gamma)$ for these centers are near 0.5 and similarly for $\widehat{P}_k$ and percentiles based on ρ^{pm}. For dialysis centers with a large number of patients and thereby a small variance, the optimally estimated percentiles, $\widehat{P}_k$ and $\tilde{P}_k(\gamma)$ spread out to cover full range from 0 to 1. There is better agreement between MLE-based, Z-score-based and optimal percentiles when the small centers are removed from the dataset and estimates are recomputed.

Figure 4 displays estimates for the 40 providers at the 1/3174, 82/3174, 163/3174,…, 3173/3174 percentiles as determined by $\widehat{P}_k$. For each display, the Y-axis is $100 \times \bar{P}_k$ with its 95% posterior interval. The X-axis for the upper left panel is $\widehat{P}$, for the upper right is percentiles based on ρ^{pm}, for the lower left is percentiles based on ρ^{mle}, and for the lower right is percentiles based on Z-scores testing $\rho_k = 1$. To deal with cases where $Y_{kt} = 0$, for the hypothesis test statistic we use

$$\text{Z}-\text{score}=\log\left(\frac{y_k}{m_k}+0.25\right)\sqrt{m_k}.$$

See Conlon and Louis (1999) for a similar plot based on SMRs of disease rates in small areas.

Note that in the upper left display the $\bar{P}_k$ do not fill out the (0, 1.0) percentile range; they are shrunken toward 0.50 by an amount that reflects estimation uncertainty. Also, the posterior probability intervals are very wide, indicating considerable uncertainty in estimating ranks/percentiles. The plotted points in the upper left display are monotone because the X-axis is the percentile based on ranking of Y-axis values. Plotted points in the upper right display, which are based on posterior mean, are almost monotone and close to the best attainable. The lower left and lower right panels 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 SMRs estimated from the pooled 1998-2001 data would be qualitatively similar, but the departures from monotonicity would be less extreme.

We divide *MSE* for different ranking methods by the *MSE* of randomly assigned ranks (Sect. 4.1) for standardization. The methods based on posterior distributions, P^{pm}, $\widehat{P}_k$, $\tilde{P}(0.8)$ and $P^{\star}(0.8)$ perform pretty close to each other with standardized *MSE*s 44.5%, 44.4%, 46.2% and 46.2%, respectively. Rankings based on MLE and Z-score have less improvement (52.3% and 47.4%) over randomly assigned ranks. The differences in $\sqrt{MSE}$ are less substantial and the wide 95% intervals presented in Fig. 4 indicate that none of methods can give a conclusive ranking result.

5.4 Single year and multi-year analyses

Using model (2) 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 (the typical log(*SMR*) = 0). The within year, between provider variation in $100 \times \log(SMR)$ is essentially constant at approximately $100 \times \tau = 24$, producing a 95% *a priori* interval for the ρ_{kt} (0.62, 1.60). While we have a prior centering around 1000 for $100 \times \tau$, the data likelihood dominates the prior information and the posterior 95% credible interval of $100 \times \tau_t$ for all 4 years is (22.8, 26.8). Use of the *AR*(1) model to combine evidence over years (with the posterior distribution for ϕ concentrated around 0.90) reduces $100 \times OC_{\tilde{p}}(0.8)$ from around 61 to around 48, a twenty percent decrease. Classification performance comparison using the $\widehat{P}_k$ is very close to that for the optimal $100 \times \tilde{P}_k(0.8)$.

Figure 1 displays the details behind the improvement of classification performance. In the upper range of $\tilde{P}_k(0.8)$, 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

$\tilde{P}_k(0.8)$, the curves would need to cross at $\tilde{P}_k(\gamma) = 0.8$, but the curves cross at about 0.7. Appendix B provides some discussion on this phenomenon.

Longitudinal variation in ranks/percentiles (LV_{Pest} 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 $\phi \to 1$, $LV_{\tilde{P}} \to 0$ and if the data provide no information on the SMRs (the $\tau \to \infty$), then $LV_{\hat{P}} = 83$.

We have not compared fit of the *AR*(1) model to other correlation structures such as compound symmetry (constant correlation rather than exponential damping). With only 4 years of data per center, the power to compare different correlation structures will be low. With ϕ's posterior mean 0.90 and 95% credible interval (0.88, 0.92), the *AR*(1) model is well supported by the data relative to independence. Note that the *AR*(1) model operates on the $\theta_{kt} = \log(\rho_{kt})$ and not on the observed estimates $\left(\theta_{kt}^{mle}\right)$. The induced model for these is approximately *ARMA*(1, 1), a hidden Markov model.

5.5 Parametric and non-parametric priors

We compare percentiles based on posterior distributions under the parametric and NPML priors using 1998 data. Figure 5 displays Gaussian, posterior expected and smoothed NPML estimated priors for $\theta = \log(\rho)$. The Gaussian is produced by plugging in the posterior medians for (μ_0, τ_0). The posterior expected is a mixture of Gaussians using the posterior distribution of (μ_0, τ_0). The posterior distribution of (μ_0, τ_0) has close to 0 variance, so the two parametric curves superimpose. The NPML is discrete and was smoothed using the "density" function in R with adjustment parameter 10 (i.e., the Gaussian kernel bandwidth is ten times of the default value, see Silverman (1986)). We smooth the NPML to graph a smooth curve, but use the NPML itself to produce ranks/percentiles. Note that the smoothed 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. (2006). Fig. 2 compares $\tilde{P}(0.8)$ under the two priors. The centers at the top or the bottom have less uncertainty in percentiles (strong signal), and their percentiles are generally same under two priors. For the dialysis centers with larger variance, the percentiles depend on the prior.

5.6 Ranks based on exceedance probabilities

We compute $P^*(0.8)$ (see Sect. 3.3) using the Gaussian prior for θ and 1998 data. The θ-threshold, $\bar{G}_K^{-1}(0.8) = 0.169$ (ρ-threshold = 1.184). Lin et al. (2006) prove the near equivalence of $P^*(\gamma)$ and $\tilde{P}_k(\gamma)$ and Fig. 1 displays this equivalence in that the curve based on $P^*(0.8)$ is nearly identical to that based on $\tilde{P}(0.8)$ for $\phi = 0$.

6 Discussion

Ranks and percentiles are computed to address specific policy or management goals. It is important to use a procedure that performs well for the primary goals. A structured approach guided by a Bayesian hierarchical model and a loss function helps clarify goals and produces ranks/percentiles that outperform other contenders, such as those based on MLEs and Z-scores. When the uncertainties of the direct estimate vary considerably over providers, the estimates are very sensitive to the method used. In that situation, a structured approach is especially important.

Our data-analytic assessments support the Lin et al. (2006) finding that the $\widehat{P}_k$ (general purpose percentiles) perform well over inferential goals addressed by a range of loss functions, but that if a specific percentile cut-point, γ, is identified, $\tilde{P}_k(\gamma)$ (or $P^*(\gamma)$) should be used. Unless the substantive application dictates otherwise, we recommend the use of these. Cost or other considerations can be incorporated to select γ.

Though the loss function guided estimates are the best possible, the ranking results might not be conclusive, partially indicated by the wide confidence interval as shown in the Fig. 4. Therefore, data-analytic performance evaluations are a necessary companion to estimated ranks. Uncertainty assessments include standard errors and tabulation or display of the probabilities of correct classification in a (above γ)/(below γ) assessment (our $\pi_k(\gamma|\mathbf{Y})$). These probabilities can be used to temper penalties or rewards. When available, data of multiple years can be combined to reduce the uncertainty in ranking results, as shown in Table 1 and Fig. 1.

Robustness of efficiency and validity are important attributes of any statistical procedure. For sufficiently large *K*, using a smoothed non-parametric prior is highly efficient relative to a correct, parametric approach and confers considerable robustness (see Paddock et al. 2006). Additional study of this strategy is needed.

Percentiles are *prima facie* relative comparisons in that it is possible that all providers are doing well or that all are doing poorly; percentiles will not pick this up. Indeed, the SMR is, itself, a relative measure and so percentiles produced from it are twice removed from a normative context! In situations where normative values are available (e.g., death rates), percentiles that have a normative interpretation are attractive and those based on posterior probabilities of exceeding some threshold ($P^*(\gamma)$) are essentially identical to a loss function based approach $\left(\tilde{P}_k(\gamma)\right)$ and so provide an excellent link to a substantively relevant scale. And, they confer a considerable computing advantage over using the posterior distribution of the ranks to find the $\tilde{P}_k(\gamma)$.

Finally, because percentiles can be very sensitive to the estimation methods and because there is considerable uncertainty associated with all percentiling methods, stakeholders need to be informed of the issues in producing percentiles, in interpreting them, in their role in science and policy, and in insisting on uncertainty assessments.

Acknowledgements

Supported by grant R01-DK61662 from U.S NIH, National Institute of Diabetes, Digestive and Kidney Diseases., We thank Chris Forrest for his advice and comments.

Appendices

Appendix A: An illustrative example of subset dependency

In general, ranks depend on the framework for comparison and the list of contenders; they are not necessarily subset invariant. For illustration, consider ranking 3 dialysis centers by $\bar{R}_k$ as defined in Eq. 3. We have,

$$\begin{aligned}\bar{R}_1 &= \mathrm{pr}(\rho_1>\rho_2)+\mathrm{pr}(\rho_1>\rho_3)\\ \bar{R}_2 &= \mathrm{pr}(\rho_2>\rho_1)+\mathrm{pr}(\rho_2>\rho_3)\\ \bar{R}_3 &= \mathrm{pr}(\rho_3>\rho_1)+\mathrm{pr}(\rho_3>\rho_2).\end{aligned}$$

Let $\theta_i=\log(\rho_i)\sim N\left(\mu_i,\sigma_i^2\right)$, i = 1, 2, 3. Let $\mu_1 = 0$, $\mu_2 = -0.15$, $\mu_3 = 0.05$, $\sigma_1^2=\sigma_3^2=0.01^2$, $\sigma_2^2=7.5^2$, then $pr(\rho_1 > \rho_2) = 0.51 > 0.49 = pr(\rho_2 > \rho_1)$; $pr(\rho_1 > \rho_2) + pr(\rho_1 > \rho_3) = 0.51 < 0.98 = pr(\rho_2 > \rho_1) + pr(\rho_2 > \rho_3)$. Thus if we rank only dialysis centers 1 and 2, center 2 has better rank (smaller θ) than center 1; if we rank all 3 centers, center 2 has worse rank than center 1.

Appendix B: Crossing points of exceedance probability curves

We start from a simplified scenario where θ_k's share a common posterior variance. Assume $\theta_k \sim N(\mu_k, \nu)$ a posteriori, $\Phi(t; \mu_k, \nu) = pr(\theta > t; \mu_k, \nu)$. Let ν_0 and ν_0' be two possible values of ν, $\nu_0 > \nu_0'$. Without loss of generality, let $\mu_1 < \mu_2 < \ldots < \mu_K$.

For any given t,

- If $t<\mu_k, \frac{t-\mu_k}{\sqrt{\nu_0}} > \frac{t-\mu_k}{\sqrt{\nu_0'}}, \Phi(t;\mu_k,\nu_0) = \Phi\left(\frac{t-\mu_k}{\sqrt{\nu_0}};0,1\right) < \Phi\left(\frac{t-\mu_k}{\sqrt{\nu_0'}};0,1\right) = \Phi\left(t;\mu_k,\nu_0'\right)$;
- If $t>\mu_k, \frac{t-\mu_k}{\sqrt{\nu_0}} < \frac{t-\mu_k}{\sqrt{\nu_0'}}, \Phi(t;\mu_k,\nu_0) = \Phi\left(\frac{t-\mu_k}{\sqrt{\nu_0}};0,1\right) > \Phi\left(\frac{t-\mu_k}{\sqrt{\nu_0'}};0,1\right) = \Phi\left(t;\mu_k,\nu_0'\right)$;

By the common variance assumption and $\mu_1 < \mu_2 < \ldots \mu_K$, the posterior distributions of θ_k are stochastically ordered and the rank of θ_k is k. The curves $\left(\frac{k}{K+1}, \Phi(t,\mu_k,\nu_0)\right)$ and $\left(\frac{k}{K+1}, \Phi\left(t,\mu_k,\nu_0'\right)\right)$ are both monotone increasing and cross each other between $\frac{i}{K+1}$ and $\frac{i+1}{K+1}$ if $\mu_i < t < \mu_{i+1}$. The value of y-coordinate of the crossing point is around 0.5 due to $\frac{t-\mu_i}{\gamma} \approx 0$. We denote the x-coordinate of the crossing point as $\frac{i(t)}{K+1}$ ignoring at most $1/K$ difference.

If t_1 and t_2 satisfy

$$\frac{1}{K}\sum_{k=1}^{K}\Phi(t_1;\mu_k,\nu_0) = \frac{1}{K}\sum_{k=1}^{K}\Phi\left(t_2;\mu_k,\nu_0'\right) = 1-\gamma$$

then the curves $\left(\frac{k}{K+1}, \Phi(t_1,\mu_k,\nu_0)\right)$ and $\left(\frac{k}{K+1}, \Phi\left(t_1,\mu_k,\nu_0'\right)\right)$ cross each other at $\frac{i(t_1)}{K+1}$; the curves $\left(\frac{k}{K+1}, \Phi(t_2,\mu_k,\nu_0)\right)$ and $\left(\frac{k}{K+1}, \Phi\left(t_2,\mu_k,\nu_0'\right)\right)$ cross each other at $\frac{i(t_2)}{K+1}$; And the crossing-over of curves $\left(\frac{k}{K+1}, \Phi(t_1,\mu_k,\nu_0)\right)$ and $\left(\frac{k}{K+1}, \Phi\left(t_2,\mu_k,\nu_0'\right)\right)$ happens between $\frac{i(t_1)}{K+1}$ and $\frac{i(t_2)}{K+1}$. When γ is greater or smaller than both of $\frac{i(t_1)}{K+1}$ and $\frac{i(t_2)}{K+1}$, which depend on γ, ν_0, ν_0' and vector (μ_1, μ_2,…, μK), the x-coordinate of the crossing point can not be at γ.

Denote the posterior distributions of θ_k as $N(\mu_k, \nu_k)$ and $N\left(\mu_k',\nu_k'\right)$ in single year and multiple years analyses. If the dialysis centers perform consistently over years, inference uncertainty of θ_k should be reduced (assuming $\nu_k' < \nu_k$) by accumulating data over years while the means do

not change much (assuming $\mu'_k = \mu_k$). Assuming $N(\mu_k, \nu_k)$'s are stochastically ordered, $N\left(\mu'_k=\mu_k,\nu'_k\right)$'s are stochastically ordered, the above discussion applies to the crossing point of the curves $\left(\frac{k}{K+1},\Phi\left(t,\mu_k,\nu_k\right)\right)$ and $\left(\frac{k}{K+1},\Phi\left(t,\mu_k,\nu'_k\right)\right)$.

In Fig. 1, two curves $\left(\frac{k}{K+1},\Phi\left(t,\mu_k,\nu_k\right)\right)$ and $\left(\frac{k}{K+1},\Phi\left(t,\mu_k,\nu'_k\right)\right)$ are plotted without the stochastically ordered assumption and the location crossing point is more complicated. In general, it is not necessary that the x-coordinate of the crossing point will be at γ.

Appendix C: The NPML algorithm

Assume $\rho_k \sim G, k = 1,\ldots, K$. G is discrete having at most J mass points $u_1,\ldots,u_J$ with probabilities $p_1,\ldots, p_J$. We use EM algorithm (Dempster et al. 1977) to estimate the u's and p's. Start with $u_1^{(0)},\ldots,u_j^{(0)}$ and $p_1^{(0)},\ldots,p_J^{(0)}$, for each recursion,

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

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 $u_j^{(\nu)}$ and the $p_j^{(\nu)}$, j = 1, 2, …,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:

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

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

$$\bar{\rho}^{(\nu)} = \sum_j p_j(\nu)\, u_j^{(\nu)},$$

and write

$$\left(m_k u_j^{(\nu)}\right)^{y_k} = \exp\left(y_k \log\left(m_k u_j^{(\nu)}\right)\right).$$

The *w*-recursion becomes:

$$w_{kj}^{(v+1)}=\frac{\left(u_j^{(v)}/\bar{\rho}^{(v)}\right)^{y_k}\exp\left(-m_k\left(u_j^{(v)}-\bar{\rho}^{(v)}\right)\right)p_j^{(v)}}{\sum_{l=1}^{J}\left(u_l^{(v)}/\bar{\rho}^{(v)}\right)^{y_k}\exp\left(-m_k\left(u_l^{(v)}-\bar{\rho}^{(v)}\right)\right)p_l^{(v)}}$$
$$=\frac{p_j^{(v)}\exp\left(y_k\log\left(u_j^{(v)}/\bar{\rho}^{(v)}\right)-m_k\left(u_j^{(v)}-\bar{\rho}^{(v)}\right)\right)}{\sum_{l=1}^{J}p_l^{(v)}\exp\left(y_k\log\left(u_l^{(v)}/\bar{\rho}^{(v)}\right)-m_k\left(u_l^{(v)}-\bar{\rho}^{(v)}\right)\right)}$$

References

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. C Stat. Methodol 1995;57:289–300.

Carlin, BP.; Louis, TA. Bayesian Methods for Data Analysis. Vol. 3rd edn.. Chapman and Hall/CRC Press; Boca Raton FL: 2008.

Christiansen C, Morris C. Improving the statistical approach to health care provider profiling. Ann. Intern. Med 1997;127:764–768. [PubMed: 9382395]

Conlon, EM.; Louis, TA. Addressing multiple goals in evaluating region-specific risk using Bayesian methods. In: Lawson, A.; Biggeri, A.; Böhning, D.; Lesaffre, E.; Viel, J-F.; Bertollini, R., editors. Disease Mapping and Risk Assessment for Public Health. Wiley; 1999. p. 31-47.chap. 3

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (C/R: P 22-37). J. R. Stat. Soc. Ser. C Stat. Methodol 1977;39:1–22.

Diggle PJ, Thomson MC, Christensen OF, Rowlingson B, Obsomer V, Gardon J, Wanji S, Takougang I, Enyong P, Kamgno J, Remme JH, Boussinesq M, Molyneux DH. Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty. Ann. Trop. Med. Parasitol 2007;101 (6):499–509. [PubMed: 17716433]

Dominici F, Parmigiani G, Wolpert RL, Hasselblad V. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. J. Am. Stat. Assoc 1999;94:16–28.

Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet. Epidemiol 2002;23:70–86. [PubMed: 12112249]

ESRD. 1999 Annual Report: ESRD Clinical Performance Measures Project. Health Care Financing Administration; 2000. Technical Report

Gelman A, Price P. All maps of parameter estimates are misleading. Stat. Med 1999;18:3221–3234. [PubMed: 10602147]

Gelman, A.; Sturtz, S.; Ligges, U.; Gorjanc, G.; Kerman, J. 2006. The R2WinBUGS package. http://cran.r-project.org/doc/packages/R2WinBUGS.pdf

Goldstein H, Spiegelhalter D. League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). J. R. Stat. Soc. Ser. A 1996;159:385–443.

Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Stat. Methods Med. Res 2003;12(2):147–170. [PubMed: 12665208]

Lacson E, Teng M, Lazarus J, Lew N, Lowrie E, Owen W. Limitations of the facility-specific standardized mortality ratio for profiling health care quality in dialysis. Am. J. Kidney Dis 2001;37:267–275. [PubMed: 11157366]

Laird NM. Nonparametric maximum likelihood estimation of a mixing distribution. J. Am. Stat. Assoc 1978;78:805–811.

Landrum M, Bronskill S, Normand S-L. Analytic methods for constructing cross-sectional profiles of health care providers. Health. Serv. Outcomes Res. Method 2000;1:23–48.

Lin R, Louis TA, Paddock SM, Ridgeway G. Loss function based ranking in two-stages, hierarchical models. Bayesian Anal 2006;1(4):915–946.

Liu J, Louis TA, Pan W, Ma J, Collins A. Methods for estimating and interpreting provider-specific, standardized mortality ratios. Health. Serv. Outcomes Res. Method 2004;4:135–149.

Lockwood J, Louis TA, McCaffrey DF. Uncertainty in rank estimation: implications for value-added modeling accountability systems. J. Edu. Behav. Stat 2002;27(3):255–270.

Louis TA, Shen W. Innovations in Bayes and empirical Bayes methods: estimating parameters, populations and ranks. Stat. Med 1999;18:2493–2505. [PubMed: 10474155]

Louis TA, Zeger SL. Effective communication of standard errors and confidence intervals. Biostatistics. 2008http://dx.doi.org/10.1093/biostatistics/kxn014

McClellan, M.; Staiger, D. The Quality of Health Care Providers. National Bureau of Economic Research; 1999. Technical Report 7327Working Paper

Normand S-LT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. J. Am. Stat. Assoc 1997;92:803–814.

Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. Stat. Sci 2007;22:206–226.

Ohlssen DI, Sharples LD, Spiegelhalter DJ. A hierarchical modelling framework for identifying unusual performance in health care providers. J. R. Stat. Soc. Ser. A Stat. Soc 2007;170(4):865–890.

Paddock S, Ridgeway G, Lin R, Louis TA. Flexible distributions for triple-goal estimates in two-stage hierarchical models. Comput. Stat. Data Anal 2006;50(11):3243–3262.

Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. Am. J. Epidemiol 2008;167(3):362–368.http://dx.doi.org/10.1093/aje/kwm305 [PubMed: 17982157]

Plummer, M.; Best, N.; Cowles, K.; Vines, K. The CODA Package. 2006.

Shen W, Louis TA. Triple-goal estimates in two-stage, hierarchical models. J. R. Stat. Soc. Ser. B 1998;60:455–471.

Shen W, Louis TA. Triple-goal estimates for disease mapping. Stat. Med 2000;19:2295–2308. [PubMed: 10960854]

Silverman, BW. Density Estimation for Statistics and Data Analysis. Chapman & Hall Ltd; 1986.

Spiegelhalter, D.; Thomas, A.; Best, N.; Gilks, W. BUGS: Bayesian Inference Using Gibbs Sampling. Vol. Version 0.60. Medical Research Council Biostatistics Unit, Institute of Public Health; Cambridge University: 1999. Technical Report

Storey JD. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Methodol 2002;64(3):479–498.

Storey JD. The positive false discovery rate: a Bayesian Interpretation and the q-value. Ann. Stat 2003;31 (6):2013–2035.

USRDS. 2005 Annual Data Report: Atlas of end-stage renal disease in the United States. Health Care Financing Administration; 2005. Technical report

Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. Int. J. Qual. Health Care 2001;13(6):481–488. [PubMed: 11769751]

Zhang M, Strawderman RL, Cowen ME, Wells MT. Bayesian inference for a two-part hierarchical model: an application to profiling providers in managed health care. J. Am. Stat. Assoc 2006;101(475):934–945.

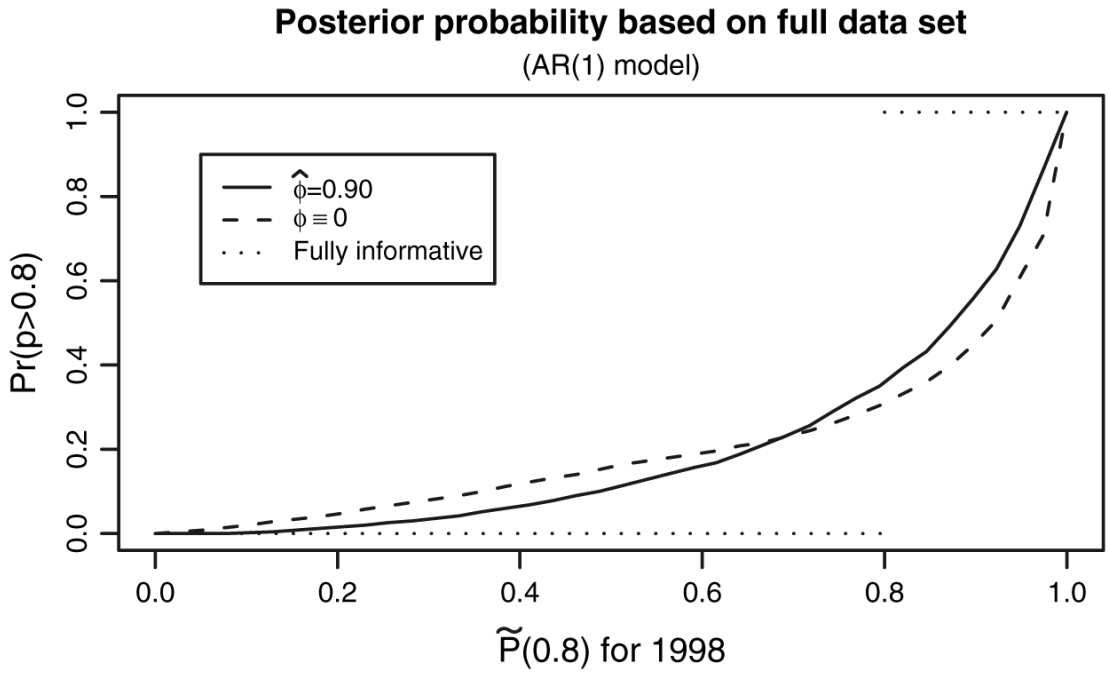

Fig. 1.

π_k (0.8|**Y**) versus $\tilde{P}_k$ (0.8) for 1998. Optimal percentiles and posterior probabilities computed with the single year model ($\phi \equiv 0$) and the *AR*(1) model ($\phi = 0.90$) Two curves don't cross at $\gamma = 0.8$. The line for fully informative data, i.e., when there is no uncertainty associated with ranking results is given as reference

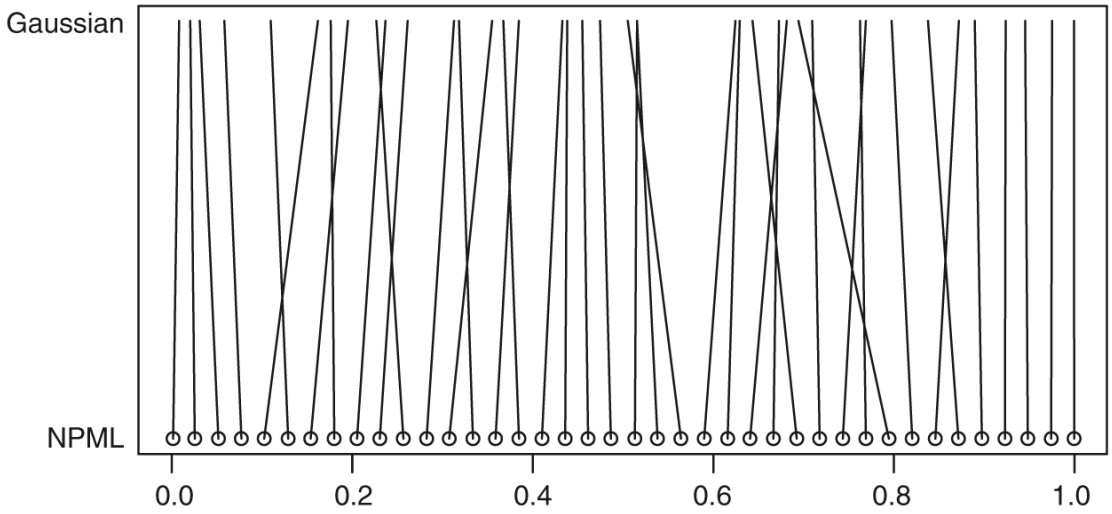

Fig. 2.

Comparison of 1998 $\tilde{P}(0.8)$ with NPML and Gaussian prior. Circles represent 40 dialysis centers evenly spread across percentiles estimated with NPML prior. The percentiles of the same center are connected

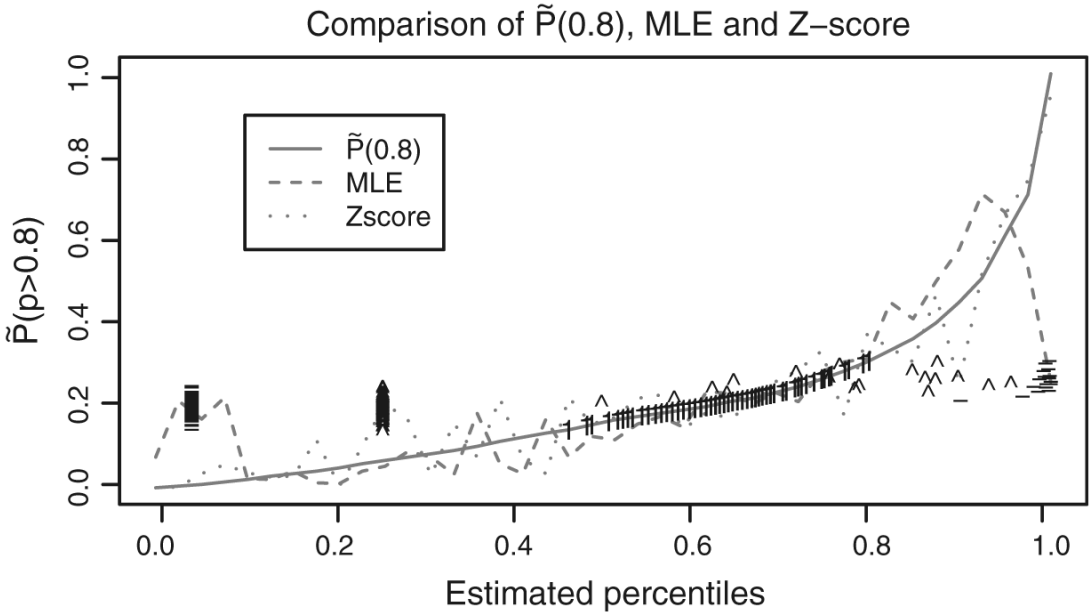

Fig. 3.

$\pi_k(0.8)$ versus estimated percentiles by three ranking methods using the 1998 data: $\tilde{P}_k(\gamma)$, MLE-based and Z-score-based. For small dialysis centers (fewer than 5 patients in 1998), the symbol "-" represents the MLE-based percentiles, the symbol "1" the Z-score-based percentiles and the symbol " ^ " the $\tilde{P}_k(\gamma)$

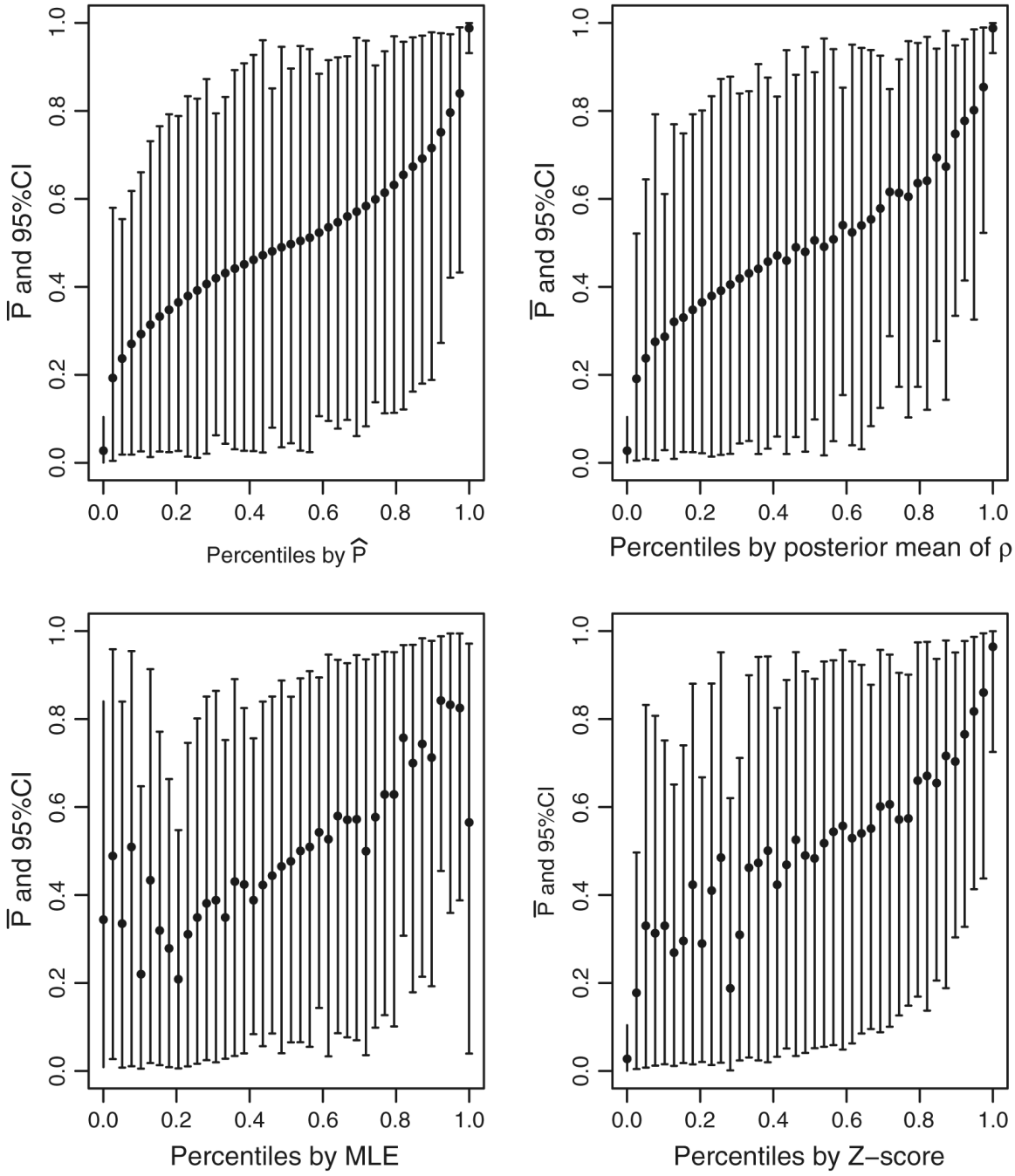

Fig. 4.

SEL-based percentiles for 1998. For each display, the Y-axis is 100 × $\bar{P}_k$ with its 95% probability interval. The X-axis for the upper left panel is $\widehat{P}$, for the upper right is percentiles based on ρ^{pm}, for the lower left is percentiles based on the ρ^{mle} and for the lower right is percentiles based on Z-scores testing $\rho_k = 1$

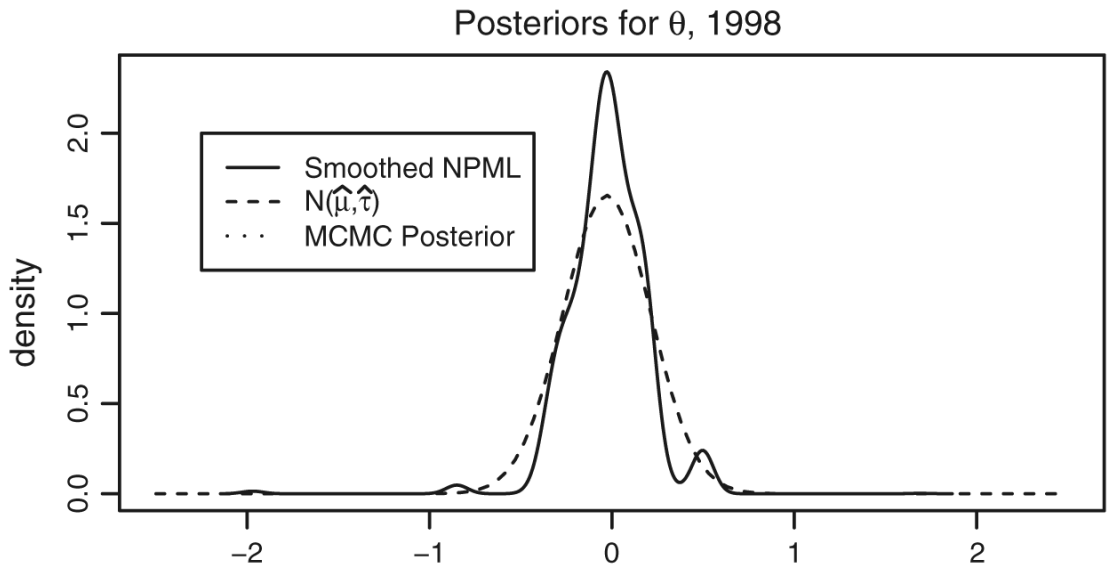

Fig. 5.

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

Table 1

Results for $\widehat{P}_k$ and $\tilde{P}(0.8)$. In the multi-year section, $100 \times OC(0.8)$ is for the indicated year as estimated from the multi-year model and $_{88}90_{92}$ is a notation for posterior median 90 and 95% credible interval (88, 92) (Louis and Zeger 2008)

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