Published in final edited form as:

Health Serv Outcomes Res Methodol. 2013 March 1; 13(1): 1–17. doi:10.1007/s10742-012-0099-5.

Assessing the accuracy of profiling methods for identifying top providers: performance of mental health care providers

Victoria Y. Ding,

Department of Biostatistics, University of Washington, Seattle, WA, USA

Rebecca A. Hubbard,

Department of Biostatistics, University of Washington, Seattle, WA, USA; Group Health Research Institute, 1730 Minor Ave., Suite 1600, Seattle, WA 98101, USA

Carolyn M. Rutter, and

Department of Biostatistics, University of Washington, Seattle, WA, USA; Group Health Research Institute, 1730 Minor Ave., Suite 1600, Seattle, WA 98101, USA

Gregory E. Simon

Group Health Research Institute, 1730 Minor Ave., Suite 1600, Seattle, WA 98101, USA

Abstract

Provider profiling as a means to describe and compare the performance of health care professionals has gained momentum in the past decade. As a key component of pay-for-performance programs profiling has been increasingly used to identify top-performing providers. However, rigorous examination of the performance of statistical methods for profiling when used to classify top-performing providers is lacking. The objective of this study was to compare the classification accuracy of three methods for identifying providers exceeding performance thresholds and to analyze data on satisfaction with mental health care providers at Group Health Cooperative using these methods. Questionnaire data on patient satisfaction with mental health care providers at Group Health Cooperative was collected between April 2008 and January 2010. Simulated data were used to compare the classification accuracy of alternative statistical methods. We evaluated sensitivity, specificity, and root mean squared error of alternative statistical methods using simulated data. For Group Health providers, we compared agreement of alternative approaches to classification. We found that when between-provider variability in performance was low, all three methods exhibited poor classification accuracy. When used to evaluate mental health care provider performance, we found substantial uncertainty in the estimates and poor agreement across methods. Based on these findings, we recommend providing uncertainty estimates for provider rankings and caution against the use of any classification method when between-provider variability is low.

Keywords

Monte Carlo methods; Pay-for-performance; Provider profiling; Random effects model; Variance components

1 Introduction

Assessing the comparative performance of health care providers is widely practiced and is a key feature of health care reform (Marshall et al. 2000). One of the primary objectives of this assessment is to evaluate the quality or value of care provided by individual practitioners, a practice referred to as “provider profiling” (Smith 2002). Information on the relative performance of health care providers may be used by individual patients to select providers, or by health care organizations to identify candidates for remedial programs or bonuses (Epstein et al. 2004; Kak et al. 2001). Moreover, policymakers may use performance data to assess whether core health care needs are being met (Hauck et al. 2003).

Provider profiling is increasingly used to identify providers exceeding performance thresholds. For instance, the majority of programs offering financial incentives for high quality care, or pay-for-performance, award incentives to providers who exceed fixed thresholds (Rosenthal and Dudley 2007). Indeed, pay-for-performance for hospital quality is mandated by the Patient Protection and Affordable Care Act of 2010. This differs from earlier uses of profiling, which focused on identifying outlying providers whose performance was far from the norm. Simple stratification into top- and bottom-performers, without considering individual outlier status, is now common and integral to health care reform. This has created a need for statistical methods that accurately classify providers as top- or bottom-performers. However, no existing research has estimated classification accuracy of statistical methods used to classify providers as top- or bottom-performers.

Several statistical methods are commonly used for provider profiling and can be adapted to the context of provider classification. For example, in the indirect standardization method, the expected score for an average provider treating patients with a given set of characteristics is estimated using a fixed-effects regression model. The total score for all a provider’s patients is then calculated and the ratio of the observed to the expected score computed. This method has been used, for example, to compare hospital-specific, riskstandardized, 30 day all-cause mortality after isolated coronary artery bypass graft surgery using data from the Society of Thoracic Surgeons National Database (Shahian and Normand 2008).

Random effects or hierarchical models (Bryk and Raudenbush 1992) are also widely used to evaluate providers using estimates of case-mix-adjusted performance (D’Errigo et al. 2007; Glance et al. 2006; Shahian et al. 2007). These models are appealing because they not only adjust for variation in case-mix, but also address differences in the precision of provider-performance estimates that arise from differences in the number of patients per provider (Burgess et al. 2000). Under this approach, fixed effects are estimated for case-mix characteristics that may be associated with the outcome, and residual between-provider variability is assessed via a provider-level random effect. Variation in sample size is addressed by borrowing information across providers to stabilize estimates for providers who have relatively few subjects. Ohlssen et al. (2007b) showed that provider performance estimates based on hierarchical models are optimal in terms of efficiency and bias.

There are also several important limitations of the random effects approach. First, several authors have cautioned against the use of random effects models because they tend to be conservative (Austin et al. 2003; DeLong et al. 1997; Ohlssen et al. 2007a; Racz and Sedransk 2010). Because of the “shrinkage” property of the random effects estimator, providers lying in the tails of the distribution for which small sample sizes are available will be substantially “shrunk” towards the population mean. This will lead to decreased sensitivity for detecting these extreme providers and overly conservative estimates. Additionally, misspecification of the random effects distribution can lead to problems in

estimation. Heagerty and Kurland (2001) found that the asymptotic relative bias of the intercept estimate in a random effects models was on the order of 30 % for highly skewed distributions and even greater when between-cluster heterogeneity was substantial. McCulloch and Neuhaus (2011) echoed this finding and additionally noted that the shape of the estimated random effects distribution will reflect the shape of the assumed distribution rather than the true underlying shape. In the profiling context, imposing a Gaussian assumption may result in attenuated estimated effects for top providers if the true distribution is more highly skewed.

Previous research has compared methods for estimating provider performance (Austin 2005; Austin et al. 2003; DeLong et al. 1997). For instance, Austin et al. (2003) compared the performance of indirect standardization and random effects models for identifying hospitals with outlying mortality rates. However, previous work has not addressed the critical question of whether these methods can be used to identify providers performing above a threshold when no outliers exist.

Our research was motivated by the use of profiling for identifying top-performing providers in the setting of mental health care. Group Health, a consumer-governed, non-profit health system in Washington state and North Idaho, has an aggressive incentive program that rewards community practitioners contracting with Group Health who score above the 90th percentile on a consolidated metric of five measures included in quarterly provider profiles (Alliance of Community Health Plans 2009). In addition, the Group Health Cooperative Behavioral Health Service Department engaged in a 5 year initiative to improve patient satisfaction; changes in the mean scores of satisfaction ratings were analyzed from 2002 to 2007 across a number of satisfaction items including access to initial appointments and coordination of care (Crosier et al. 2012). However, a study on depression outcomes for primary care providers by Katon et al. (2000) found no important differences in quality of care or patient outcomes, when random effects models were used to estimate mental health care provider quality. Therefore, to help evaluate the use of provider profiling at Group Health we compared the classification accuracy of three methods for ranking providers as top- or bottom-performers using data from Group Health on satisfaction with mental health care providers.

This report consists of two parts. We first compared the classification accuracy of three statistical methods for identifying top performing providers using simulated data. By using simulated data we were able to assess the performance of statistical methods in a setting in which true provider ranks were known. We then applied these statistical methods to data from Group Health, using the results of simulation studies to guide our interpretation of results, and compared the performance of alternative methods.

2 Methods

2.1 Notation and definition of classification methods

In estimation of provider performance, we assume there are N providers, with the i th provider having n_i patients and the j th patient of the i th provider having outcome measure Y_{ij} . We focus on binary outcome measures, although similar considerations apply to continuous measures. While a provider's true performance is independent of patient characteristics, these characteristics may affect patient outcomes and therefore observed performance. X_{ij} is a vector of patient characteristics, such as age, gender, or disease severity; and θ_i is the i th provider's true performance score, with the assumption that θ_i arises from a possibly unknown distribution with mean μ and variance σ^2 .

In a sample of providers, the provider with the poorest performance will have the smallest θ . The true rank of the i th provider, based on his/her performance, θ_i is given by

$$R(\theta_i) = \sum_{j=1}^N 1(\theta_i \geq \theta_j),$$

where $1(\cdot)$ is an indicator function that takes the value 1 if its conditions are satisfied and 0 if not. We estimate the rank of the i th provider by

$$R(\theta_i) = \sum_{j=1}^N 1(\widehat{\theta}_i \geq \widehat{\theta}_j),$$

where $\widehat{\theta}_i$ is the estimated performance of the i th provider. Below we discuss three methods for estimating $\widehat{\theta}_i$.

2.2 Methods for estimating provider performance

Indirect standardization is a simple method that adjusts the provider-specific proportion of subjects with a positive response based on measured patient characteristics. A standardized provider performance estimate is calculated as the ratio of the sum of observed outcomes to the sum of expected outcomes had the patients been treated by an average provider in the reference population. This is accomplished via a two-step estimation procedure. First, observations from all providers are used to estimate a fixed-effects logistic regression model relating patient characteristics to the outcome:

$$\log(p_{ij} / (1 - p_{ij})) = X_{ij}\beta,$$

where Y_{ij} is assumed to be Bernoulli distributed with mean p_{ij} . Provider performance is then estimated as

$$\widehat{\theta}_i^{(I)} = \sum_{j=1}^{n_i} Y_{ij} / \sum_{j=1}^{n_i} \widehat{p}_{ij},$$

where $\widehat{p}_{ij}$ is the fitted probability for the j th patient of the i th provider from the fixed effects logistic regression model. This estimate is a ratio of the observed number of positive responses to the expected number of positive responses for provider i .

The *fixed effects* model is an alternative method to obtain risk-adjusted estimates of provider performance. A fixed effects model takes the form

$$\log(p_{ij}^F / (1 - p_{ij}^F)) = X_{ij}\beta^F + \alpha_i^F,$$

where Y_{ij} is assumed to be Bernoulli distributed with mean p_{ij}^F , which is assumed to depend on patient characteristics and provider. Provider performance is assessed via the provider-specific parameters, $\widehat{\theta}_i^{(F)} = \widehat{\alpha}_i^F$, which are estimated as fixed-effects.

Finally, a *random effects* model estimates provider performance assuming common effects of patient characteristics across providers and allowing for between-provider variability in the outcome via a random effect. We used a random intercepts model of the form

$$\log(p_{ij}^R / (1 - p_{ij}^R)) = X_{ij}\beta^R + \alpha_i^R,$$

where Y_{ij} is assumed to be Bernoulli distributed with mean p_{ij}^R , again assumed to depend on both patient characteristics and provider. However, here the provider effects—the random intercepts, α_i —are assumed to arise from a common distribution with mean μ and variance σ^2 . In this method, provider performance is assessed using the estimated provider-specific random effects, $\hat{\theta}_i^{(R)} = \hat{\alpha}_i^{(R)}$, with higher values corresponding to better performance.

2.3 Assessing classification performance

Top- and bottom-performing providers are sometimes identified for bonuses or remedial programs. If each provider's true performance were known, then we could classify a subset of providers into a top-performing group based on rank. However, true performance is generally unknown, so estimates such as those described above are used to classify providers.

To identify $(100 \times p_0)\%$ of N providers as “top” performers, we assigned ranks to providers based on their performance estimates. Top performers were defined as those with ranks $N \times (100 \times p_0) + 1$ to N . The classification performance of an estimator can be assessed in a variety of ways. In this study, we focused on two measures: sensitivity, defined as the proportion of top-performing providers that are correctly classified; and specificity, defined as the proportion of non-top-performers that are correctly classified.

2.4 Simulation study design

Since true provider performance is unobservable, classification performance was assessed using simulated data. We simulated data representing N providers using Poisson-distributed patient volumes. Values for N , mean patient volume, and between-provider variability were motivated by data on satisfaction with mental health care at Group Health (see description below). We simulated each provider's true performance, α_j , first from a normal distribution with variance σ^2 and then from a bimodal distribution arising from a mixture of normals, with performance for 80 % of providers arising from a normal distribution with mean 0 and variance σ^2 and 20 % from a normal distribution with mean 3σ and variance σ^2 . This set of scenarios represents a setting where a subset of providers have truly extreme performance compared to the majority. We explored a range of σ^2 values to assess the impact of between-provider variability on classification accuracy. As a measure of patient variation, we simulated a patient-level covariate ranging from -5 to 5 for the j th patient of the i th provider from a truncated normal distribution. This covariate represents a patient characteristic that is associated with his/her perception of the provider's performance. Finally, we simulated binary performance outcomes for each patient from a Bernoulli distribution with mean

$$q_{ij} = \exp(\alpha_i + X_{ij}\gamma) / (1 + \exp(\alpha_i + X_{ij}\gamma)),$$

where α_i is the true performance score for the i th provider, X_{ij} is the covariate value for the j th patient of the i th provider, and $\exp(\gamma)$ is the odds ratio (OR) associated with a one unit difference in X_{ij} . Note that when $\gamma = 0$ (OR = 1), patient characteristics have no effect on observed performance and no adjustment for case-mix variation is necessary.

For each simulated data set, we estimated performance using each method, classified providers to determine the top 20 % for each method, and evaluated classification performance using sensitivity and specificity. We also evaluated the bias and variance of the estimated provider ranks using the root mean squared error (RMSE). All results are based on 1000 replications. We assessed the Monte Carlo standard error of our simulation studies using the approach recommended by Koehler et al. (2009), which yielded a maximum Monte Carlo standard error of 0.004 for sensitivity and specificity across all simulation scenarios. Finally, we simulated a data set of comparable size and variability to survey data from Group Health mental health care providers (described below) to summarize agreement among the three methods. We used R 2.10.1 (R Development Core Team, Vienna, Austria) for simulations.

2.5 Description of Group Health provider satisfaction data

We investigated the performance of classification methods using data from mailed consumer satisfaction surveys used to evaluate Group Health providers. Group Health is a not-for-profit prepaid health plan serving approximately 600,000 members. The Group Health enrollment is similar to the area population in income, educational attainment, and racial and ethnic group representation. Group Health provides mental health services using both a salaried staff of group-model providers and a contracted network of external fee-for-service providers. Group-model providers primarily serve members living in or near the cities of Seattle, Bellevue, Tacoma, Olympia, and Spokane. Network providers serve members living in or near Everett, Bellingham, and Richland as well as members living in more rural areas. Group Health conducts routine satisfaction surveys of members making individual visits to group or network mental health providers.

We included surveys returned between April 2008 and January 2010 from patients who were 18 years or older at the time of their visit. We excluded providers with fewer than five surveys. Our analysis focused on a single item, “How would you rate how well this practitioner understood your concerns?” We chose to focus on this item because it has been used by Group Health to determine which providers receive bonuses or are recommended for additional training. Providers were rated on a 5-point scale with 5 = excellent and 1 = poor. As is typical for satisfaction surveys (Lebow 1982), responses were skewed toward the positive end of the scale, with 92.9 % of responses being 4 or 5. We thus dichotomized the categorical outcome into satisfied, defined as a rating of excellent, or unsatisfied, defined as any other rating.

The survey response rate was 34 %, and provider rankings in an earlier sample did not appear to be biased by differential response rates among patient subgroups (Simon et al. 2009). All procedures were reviewed and approved by Group Health’s Human Subjects Review Committee. Consistent with applicable regulations, the committee granted a waiver of consent for research use of deidentified data from the satisfaction survey and computerized records.

We used the three methods described above to estimate risk-adjusted provider performance and rank providers from Group Health. We also quantified the uncertainty of the estimated provider ranks using a Bayesian estimation method. Diffuse prior distributions were assumed for the parameters in each model. Specifically, for classification via the fixed effects regression model, each provider’s risk-adjusted mean was assumed to be normally distributed, centered at 0 with variance 10. All other fixed effects in logistic regression models were assumed to be normally distributed, with mean 0 and variance 1,000. A more informative prior was used for the provider-specific fixed effects due to the relatively small sample sizes per provider, which necessitated more informative priors in order to achieve convergence. In the random effects model, each provider’s random effect was assumed to be

normally distributed, with mean 0 and precision following a gamma distribution with mean 1 and variance 100.

We then used a Markov chain Monte Carlo simulation to generate 1,000 samples from the posterior distribution of the provider ranks. In all cases, we used a burn-in phase of 5,000 iterations. After burn-in, 5,000 iterations were run for all three methods, and then thinned by a factor of 5. Convergence was first assessed visually for each parameter via trace plots and then via the Heidelberger and Welch diagnostic (1983). A 95 % credible interval for the ranks was constructed based on this sample. WinBUGS 1.4.3 was used for Bayesian estimation. R 2.10.1 was used for all other data analyses.

3 Results

3.1 Simulation study

3.1.1 Comparison of methods—We investigated the relationship between classification performance and degree of between-provider variability and strength of case-mix effects for simulated samples of 50 providers with a mean of 50 patients per provider and provider performance arising from a normal distribution (Table 1) or a bimodal distribution arising from a mixture of normals (Table 2). The three methods investigated performed similarly across most scenarios. Under scenarios with no effect of case-mix variation ($OR = 1$) all three methods performed almost identically. As the effect of case-mix variation increased, the indirect standardization method demonstrated somewhat poorer sensitivity compared to other methods. For instance, under the normal distribution with $\sigma^2 = 1$ and $OR = 1.5$ sensitivity for the indirect standardization method was 71 % compared to 79 % for the random effects method. For increasing case-mix effects and between-provider variability, the indirect standardization method also demonstrated increased RMSE compared to the fixed and random effects methods under both distributions for provider performance.

3.1.2 Effects of between-provider variability—The degree of between-provider variability had a substantial effect on performance. Under the lowest level of between-provider performance investigated ($\sigma^2 = 0.01$) sensitivity was below 50 % for all methods under both distributions. RMSE of ranks was also very high at this low level of variability ($RMSE > 15$) indicating substantial bias and uncertainty in rank estimates. Sensitivity and specificity increased and RMSE decreased with increasing between-provider variability. However, even at the highest level of between-provider variability investigated ($\sigma^2 = 1$) sensitivity was below 80 % for all methods under the normal distribution and below 90 % under the bimodal distribution.

3.1.3 Effects of strength of case-mix effect—Compared to the importance of between-provider variability, the strength of case-mix effects had relatively little effect on performance. For instance, under the bimodal distribution with $\sigma^2 = 1$, increasing the case-mix OR from 1 to 2 resulted in a decrease in sensitivity for the indirect standardization method from 85 to 71 %. Sensitivity of the random effects method decreased from 89 to 86 %. Changes in performance for increasing effects of case-mix variation were largest for the indirect standardization method and weaker for other methods. RMSE also increased as the effects of case-mix variation increased, and this effect was strongest for indirect standardization and weaker for the other methods considered.

3.1.4 Comparison of performance under normal and bimodal distributions—All three methods had higher sensitivity and specificity and lower RMSE under the bimodal distribution compared to the normal distribution. The random effects method performed equally as well as or better than the fixed effects method even under the bimodal distribution

where the assumed random effects distribution was misspecified. Trends observed with respect to the effect of between-provider variability and effect of case-mix variation were similar across the two distributions.

3.1.5 Effects of patient volume—We explored the dependence of classification performance on patient volume, assuming an OR of 1.5 for the effect of case-mix variation and $\sigma^2 = 0.1$, for mean patient volumes ranging from 20 to 100 patients per provider. All three methods showed improvements in sensitivity and specificity as sample size increased under both the normal and bimodal distributions for provider random effects (Fig. 1). The fixed and random effects methods had slightly higher sensitivity and specificity compared to the indirect standardization method for all patient volumes. Specificity of the random and fixed effects methods was identical for most patient volumes. For small patient volumes sensitivity was relatively low for all three methods under both distributions.

3.1.6 Comparison of ranks—Finally, we evaluated agreement of classification among the three methods using simulated data for 55 providers with true performance values arising from a normal distribution with moderate between provider variability ($\sigma^2 = 0.2$), an average of 30 patients per provider, and an odds ratio of 1.5 for the effect of case-mix variation. These values were selected to reflect a setting similar to the Group Health context. Ranks from all methods agreed closely (Fig. 2). However, despite the strength of agreement, there was substantial misclassification of the true top and bottom 20 % of providers. Indeed, one top-performer was erroneously classified as a bottom-performer. This is consistent with the low sensitivity and specificity seen for moderate between-provider variability in Table 1.

3.2 Analysis of satisfaction with Group Health mental health care providers

We investigated the performance of the three classification methods using data on satisfaction with mental health care from Group Health. Our sample consisted of 1,742 surveys for providers in the Group Health integrated practice (group) and 1,522 surveys for providers in the community network (network). These surveys represented responses for 55 group providers and 128 network providers. For group providers, the number of surveys per provider ranged from 9 to 69 with a median of 30. For network providers, the range was 5 to 33 with a median of 10. Comparison between the two groups might be biased by the selection of different patients into care with group and network providers (Simon and Ludman 2010).

Using adjusted random effects models, we estimated the association between providing an “excellent” satisfaction survey rating and risk factors: patient age (categorized as 18–24, 25–39, 40–64, and 65 or above), sex, a binary indicator of enrollment in Group Health for at least 24 months, and a binary indicator for whether the surveyed visit was a return or initial psychotherapy visit (Table 3). Choice of risk factors was based on a previous study of patient characteristics associated with satisfaction with psychotherapy (Simon et al. 2009). In the group sample, no risk factor was significantly associated with satisfaction. In the network sample, only age was statistically significantly associated with satisfaction ($p < 0.001$).

We hypothesized that there would be less variability between group providers, who practice in a common setting and participate in common training, than among network providers. Based on the adjusted random effects model, the standard deviation of the random effects for group model providers was 0.207 and for network providers was 0.548. We used estimated fixed effects for risk factors and provider-specific random effects from this model to estimate the expected proportion of patients returning “excellent” satisfaction ratings for each provider and each combination of patient characteristics. We then took the weighted

average of these estimates for each provider, weighted by the observed frequency of patient characteristics across the entire sample, to obtain a standardized estimate of the expected proportion of patients rating their satisfaction as “excellent” for each provider. In the group model, these estimates ranged from 56.2 % for the poorest performing provider to 67.5 % for the top performer and from 41.5 to 80.9 % for network providers. The difference between the top 20 % and bottom 20 % of group providers was 3.8 %, while the top and bottom network providers differed by 12.4 %.

We estimated provider performance for the two samples using each of the three methods. We adjusted all methods for the risk factors described above. Because no gold standard performance measure was available, we compared provider rankings based on the three methods. Figure 3 shows ranks for group providers based on performance scores estimated using each of the three methods. Ranks based on all three methods agreed well, with the fixed effects and indirect standardization methods agreeing extremely closely. Similar results were found for network providers (Fig. 4), despite the greater degree of variability among the network providers and the presence of significant effects of case-mix variation.

We plotted 95 % credible intervals for the estimated ranks for group providers (Fig. 5). For all three methods, intervals for most top performers extend substantially below the 80th percentile while those for most bottom performers extend substantially above the 20th percentile. Estimates based on the random effects approach exhibited greater variation than those based on the fixed effects or indirect standardization approaches. Similar results were found for network providers (Fig. 6).

4 Conclusions

In this paper, we compared the sensitivity and specificity of three statistical methods for identifying top- and bottom-performing providers—an indirect standardization method, a fixed effects model, and a random effects model. Our simulations examined a range of values for patient volume, number of providers, variation in patients, strength of case-mix effects, and between-provider variability, as well as two distributions for true provider performance. The degree of between-provider variability was the dominant factor influencing performance of methods for classifying providers. The three methods investigated performed similarly, with the random effects method very slightly outperforming other methods, even under misspecification of the random effects distribution. Indirect standardization had the poorest sensitivity, specificity and RMSE of the three methods investigated and was influenced by effects of case-mix variation more than other methods. However, in general, all methods performed poorly when there was little between provider variability. Adequate performance was observed (sensitivity near 90 % and specificity over 95 %) only for the bimodal distribution under the highest level of between-provider variability.

Our comparison of classification performance when true provider performance arose from a normal distribution or a bimodal distribution arising from a mixture of normals indicated better classification accuracy and lower RMSE for all three methods under the bimodal distribution. The random effects method performed very similarly to, and in some cases slightly better than, the fixed effects method even under the bimodal distribution where the assumed random effects distribution was misspecified. The improved performance of all methods under the bimodal distribution is likely attributable to the fact that, in this setting, a subset of 20 % of providers had substantially differing performance and hence was easier to identify. However, under the bimodal distribution the marginal variance of provider performance was also larger. We observed classification accuracy and precision of ranks to

be strongly influence by between-provider variability. Thus, in this setting of increased variability, it is not surprising that we observed improved performance.

Our analysis of Group Health provider performance indicated relatively little variability between providers both within the Group Health practice and in the contracted community network. We compared classifications made by the three methods and found high levels of agreement across methods, consistent with our simulation studies. This comparison highlights the fact that agreement among alternative statistical methods is not indicative of the accuracy of estimates. Based on our simulation study, we anticipate that the three methods provide highly consistent but erroneous classifications in settings with low between-provider variability. Based on these findings, we urge caution in interpreting agreement in classification across methods as an indication of accurate classification.

Estimates of variability in provider ranks indicated substantial uncertainty. In many cases credible intervals extended well below the 80th percentile for “top” performers and well above the 20th percentile for “bottom” performers. This problem was particularly severe for the random effects approach. Given the high degree of uncertainty in this setting, it is not surprising that misclassification is common. Previous studies have emphasized the importance of incorporating uncertainty estimates when reporting profiling results (Christiansen and Morris 1997). An alternative to classifying providers based on estimated ranks is to require some minimum probability of exceeding a performance threshold based on the posterior distribution of the estimates (Austin 2008; Christiansen and Morris 1997). At a minimum, we advise estimating standard errors, confidence intervals, or credible intervals for ranks as a means of assessing whether or not classification is reliable in a given setting.

Classification of providers based on performance scores is likely to become increasingly common as part of pay-for-performance programs. Several reviews have indicated moderate effectiveness of pay-for-performance programs (Petersen et al. 2006; Rosenthal and Frank 2006). However, the optimal design of such programs has not yet been determined. Our findings highlight the importance of considering the degree of between-provider variability when designing a pay-for-performance program. Past investigations of profiling methods for outlier detection noted the importance of appropriately accounting for differences in patient volumes and characteristics across providers (Austin et al. 2003; Austin et al. 2001; Burgess et al. 2000; Cohen et al. 2009; DeLong et al. 1997; Ohlssen et al. 2007b; Racz and Sedransk 2010). In the context of classification, we found that these characteristics also play a role. However, between-provider variability was the predominant factor influencing classification accuracy.

Overall, we found that the statistical methods included in this study are not appropriate for classifying providers as top- or bottom-performers when there is little between-provider variability. This is intuitive: if differences between providers are minimal, then accurately identifying top- and bottom performers will be difficult, and classification might be meaningless (Lockwood et al. 2002). We conclude that exploration of between provider variability is an important first step in any provider profiling analysis.

Acknowledgments

This research was supported by National Institute of Mental Health grant P20 MH068572.

References

- [Accessed 22 Jan 2012] Alliance of Community Health Plans: Quality incentives for network providers. 2009. <http://www.achp.org/permalink/1213.html><http://www.achp.org/permalink/1213.html>
- Austin PC. The reliability and validity of Bayesian measures for hospital profiling: a Monte Carlo assessment. *J. Stat. Plan. Inference.* 2005; 128(1):109–122.
- Austin PC. Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals. *BMC Med. Res. Methodol.* 2008; 8:30. [PubMed: 18474094]
- Austin PC, Naylor CD, Tu JV. A comparison of a Bayesian vs. a frequentist method for profiling hospital performance. *J. Eval. Clin. Pract.* 2001; 7(1):35–45. [PubMed: 11240838]
- Austin PC, Alter DA, Tu JV. The use of fixed- and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. *Med. Decis. Making.* 2003; 23(6):526–539. [PubMed: 14672113]
- Bryk, AS.; Raudenbush, SW. Hierarchical linear models. Sage; Newbury Park: 1992.
- Burgess JF, Christiansen CL, Michalak SE, Morris CN. Medical profiling: improving standards and risk adjustments using hierarchical models. *J. Health Econ.* 2000; 19(3):291–309. [PubMed: 10977193]
- Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann. Intern. Med.* 1997; 127(8 Pt 2):764–768. [PubMed: 9382395]
- Cohen ME, Dimick JB, Bilimoria KY, Ko CY, Richards K, Hall BL. Risk adjustment in the American College of Surgeons National Surgical Quality Improvement Program: a comparison of logistic versus hierarchical modeling. *J. Am. Coll. Surg.* 2009; 209(6):687–693. [PubMed: 19959035]
- Crosier M, Scott J, Steinfeld B. Improving satisfaction in patients receiving mental health care: a case study. *J. Behav. Health Serv. Res.* 2012; 39(1):42–54. [PubMed: 21847711]
- D'Errigo P, Tosti ME, Fusco D, Perucci CA, Seccareccia F. Use of hierarchical models to evaluate performance of cardiac surgery centres in the Italian CABG outcome study. *BMC Med. Res. Methodol.* 2007; 7:29. [PubMed: 17608921]
- DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. *Stat. Med.* 1997; 16(23):2645–2664. [PubMed: 9421867]
- Epstein AM, Lee TH, Hamel MB. Paying physicians for high-quality care. *N. Engl. J. Med.* 2004; 350(4):406–410. [PubMed: 14736934]
- Glance LG, Dick AW, Osler TM, Mukamel DB. Accuracy of hospital report cards based on administrative data. *Health Serv. Res.* 2006; 41(4):1413–1437. [PubMed: 16899015]
- Hauck K, Rice N, Smith P. The influence of health care organisations on health system performance. *J. Health Serv. Res. Policy.* 2003; 8(2):68–74. [PubMed: 12820667]
- Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika.* 2001; 88(4):973–985.
- Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. *Oper. Res.* 1983; 31(6):1109–1144.
- Kak N, Burkhalter B, Cooper M-A. Measuring the competency of healthcare providers. *Qual. Assur. Proj.* 2001; 2(1):1–28.
- Katon W, Rutter CM, Lin E, Simon G, Von Korff M, Bush T, Walker E, Ludman E. Are there detectable differences in quality of care or outcome of depression across primary care providers? *Med. Care.* 2000; 38(6):552–561. [PubMed: 10843308]
- Koehler E, Brown E, Haneuse SJ. On the assessment of Monte Carlo error in simulation-based statistical analyses. *Am. Stat.* 2009; 63(2):155–162. [PubMed: 22544972]
- Lebow JL. Consumer satisfaction with mental health treatment. *Psychol. Bull.* 1982; 91(2):244–259. [PubMed: 7071260]
- Lockwood JR, Louis TA, McCaffrey DF. Uncertainty in rank estimation: implications for value-added modeling accountability systems. *J. Educ. Behav. Stat.* 2002; 27(3):255–270. [PubMed: 19830272]

- Marshall MN, Shekelle PG, Leatherman S, Brook RH. The public release of performance data: what do we expect to gain? A review of the evidence. *JAMA*. 2000; 283(14):1866–1874. [PubMed: 10770149]
- McCulloch CE, Neuhaus JM. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*. 2011; 67(1):270–279. [PubMed: 20528860]
- Ohlssen DI, Sharples LD, Spiegelhalter DJ. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Stat. Med.* 2007a; 26(9):2088–2112. [PubMed: 16906554]
- 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.* 2007b; 170(4):865–890.
- Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? *Ann. Intern. Med.* 2006; 145(4):265–272. [PubMed: 16908917]
- Racz MJ, Sedransk J. Bayesian and frequentist methods for provider profiling using risk-adjusted assessments of medical outcomes. *J. Am. Stat. Assoc.* 2010; 105(489):48–58.
- Rosenthal MB, Dudley RA. Pay-for-performance: will the latest payment trend improve care? *JAMA*. 2007; 297(7):740–744. [PubMed: 17312294]
- Rosenthal MB, Frank RG. What is the empirical basis for paying for quality in health care? *Med. Care Res. Rev.* 2006; 63(2):135–157. [PubMed: 16595409]
- Shahian DM, Normand S-LT. Comparison of “risk-adjusted” hospital outcomes. *Circulation*. 2008; 117(15):1955–1963. [PubMed: 18391106]
- Shahian DM, Silverstein T, Lovett AF, Wolf RE, Normand S-LT. Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation*. 2007; 115(12):1518–1527. [PubMed: 17353447]
- Simon G, Rutter C, Crosier M, Scott J, Operskalski BH, Ludman E. Are comparisons of consumer satisfaction with providers biased by nonresponse or case-mix differences? *Psychiatr. Serv.* 2009; 60(1):67–73. [PubMed: 19114573]
- Simon GE, Ludman EJ. Predictors of early dropout from psychotherapy for depression in community practice. *Psychiatr. Serv.* 2010; 61(7):684–689. [PubMed: 20592003]
- Smith PC. Measuring health system performance. *Eur. J. Health Econ.* 2002; 60(3):145–148. [PubMed: 15609138]

Fig. 1.

Comparison of estimated sensitivity and specificity for three methods of classification into top 20 % of providers for provider effects simulated from normal (*left*) and bimodal (*right*) distributions. *IS* indirect standardization, *FE* fixed effects, *RE* random effects. Data are for 50 providers with values of mean patient volume (cluster size) varying from 20 to 100; effect of patient variation $OR = 1.5$; degree of between-provider variability $\sigma^2 = 0.1$

Fig. 2.

Agreement of ranks of 55 providers based on three methods for estimating provider performance in simulated data with provider effects from a normal distribution. $OR = 1.5$, $\sigma^2 = 0.2$, mean patient volume = 30. *Triangles* indicate true top 20 %, *empty circles* indicate true middle 60 %, and *filled circles* indicate true bottom 20 %. *Horizontal and vertical lines* represent cutpoints for the 20th and 80th percentiles of ranks

Fig. 3.

Agreement of ranks of 55 group model providers based on three methods for estimating provider performance. *Horizontal* and *vertical lines* represent cutpoints for the 20th and 80th percentiles of providers

Fig. 4.

Agreement of ranks of 128 network providers based on three methods for estimating provider performance. *Horizontal* and *vertical lines* represent cutpoints for the 20th and 80th percentiles of providers

Fig. 5.

95 % credible intervals for ranks of 55 group model providers based on three methods for estimating provider performance. Providers are indexed in decreasing order of estimated performance, as determined by the posterior rank. Intervals for the estimated *top* and *bottom* 20% of providers are displayed in *black*. *Horizontal lines* represent 80th and 20th percentiles of providers

Fig. 6. 95 % credible intervals for ranks of 128 network providers based on three methods for estimating provider performance. Providers are indexed in decreasing order of estimated performance, as determined by the posterior rank. Intervals for the estimated *top* and *bottom* 20 % of providers are displayed in *black*. *Horizontal lines* represent 80th and 20th percentiles of providers

Estimated sensitivity, specificity, and root mean squared error (RMSE) of provider ranks for three methods of classification into top 20 % of providers when provider effects are simulated from a normal distribution

Table 1

		Odds ratio ^a									
		1		1.5		2					
σ^2 ^b											
		0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	1
Sensitivity											
Indirect standardization		0.34	0.57	0.78	0.31	0.54	0.71	0.31	0.50	0.68	
Fixed effects		0.34	0.57	0.79	0.32	0.55	0.78	0.30	0.52	0.76	
Random effects		0.34	0.58	0.80	0.32	0.56	0.79	0.31	0.53	0.77	
Specificity											
Indirect standardization		0.83	0.89	0.94	0.83	0.88	0.93	0.83	0.87	0.92	
Fixed effects		0.83	0.89	0.95	0.83	0.89	0.94	0.83	0.88	0.94	
Random effects		0.84	0.89	0.95	0.83	0.89	0.95	0.83	0.88	0.94	
RMSE											
Indirect standardization		16.85	11.03	5.04	17.37	11.73	6.16	17.69	12.55	7.27	
Fixed effects		16.96	11.21	4.99	17.41	11.71	5.31	17.69	12.39	6.00	
Random effects		16.83	11.00	4.86	17.28	11.47	5.18	17.54	12.1	5.79	

^aOdds ratio for patient characteristics
^bBetween-provider variance

Estimated sensitivity, specificity, and root mean squared error (RMSE) of provider ranks for three methods of classification into top 20 % of providers when provider effects are simulated from a bimodal distribution

Table 2

	Odds ratio ^a									
	1		1.5		2					
	σ^2 ^b									
	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	1
Sensitivity										
Indirect standardization	0.46	0.75	0.85	0.42	0.69	0.74	0.41	0.65	0.71	
Fixed effects	0.44	0.74	0.88	0.42	0.72	0.87	0.40	0.69	0.86	
Random effects	0.46	0.76	0.89	0.44	0.74	0.88	0.42	0.71	0.86	
Specificity										
Indirect standardization	0.86	0.94	0.96	0.86	0.92	0.94	0.85	0.91	0.93	
Fixed effects	0.86	0.94	0.97	0.86	0.93	0.97	0.85	0.92	0.96	
Random effects	0.86	0.94	0.97	0.86	0.93	0.97	0.86	0.93	0.97	
RMSE										
Indirect standardization	15.25	8.72	4.27	15.74	9.66	5.66	16.18	10.59	6.86	
Fixed effects	15.49	8.93	4.08	15.90	9.49	4.32	16.28	10.22	4.88	
Random effects	15.24	8.68	3.93	15.64	9.19	4.18	15.98	9.87	4.65	

^aOdds ratio for patient characteristics

^bBetween-provider variance

Table 3

Estimated odds ratios (OR), 95 % confidence intervals (CI), and *p* values for association between “excellent” satisfaction ratings and patient characteristics in group model and network settings

	Group			Network		
	OR	95 % CI	<i>p</i>	OR	95 % CI	<i>p</i>
Age (ref > 24 years)			0.828			<0.001
25–39 years	0.96	(0.60, 1.52)		1.31	(0.83, 2.07)	
40–64 years	1.05	(0.68, 1.63)		1.64	(1.08, 2.48)	
65 years	1.07	(0.65, 1.75)		1.21	(0.70, 2.08)	
Sex (ref = female)						
Male	0.84	(0.67, 1.04)	0.106	0.98	(0.76, 1.28)	0.908
Initial or return visit (ref = initial)						
Return	1.22	(1.00, 1.5)	0.055	1.21	(0.89, 1.64)	0.217
Enrollment in GH (ref = 24 months)						
24 months	0.96	(0.77, 1.2)	0.732	0.99	(0.77, 1.27)	0.931
GH Group Health						