Assessing the Accuracy of Provider Profiling Methods for Classification

Victoria Ding

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2012

Reading Committee:

Rebecca Hubbard, Chair

Carolyn Rutter

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Assessing the Accuracy of
Provider Profiling Methods for Classification

Victoria Ding

Chair of the Supervisory Committee:
Dr. Rebecca Hubbard
Biostatistics

Provider profiling as a means to describe and compare performance of health care professionals has gained great momentum in the past decade. The implications of profiling, which can drive provider incentives and guide health policy, call for precise and accurate statistical methods. We used a simulation study to compare the performance of three commonly used methods for estimating provider performance (ranking) and for identifying high performing providers (classifying). We evaluated classification performance based on sensitivity and specificity and ranking performance based on mean squared error. We found that when between-provider variability in performance was low, all three methods performed poorly, with low accuracy for identifying top performers and high mean squared error for ranking. We then demonstrated the performance of these methods in an application to data on satisfaction with mental health care providers. Based on these findings, we caution against the use of any classification method in the setting of low between-provider variability and recommend the use of risk-adjusted methods, which take into account variation in characteristics of providers' patients, when the ratio of between-provider variability to within-provider variability is high.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to thank members of [B-faculty]—particularly Rebecca Hubbard—and [B-students] for sharing their knowledge and, of equal importance, good laughs.

DEDICATION

to my constants in times of change

Chapter 1

INTRODUCTION

This thesis explores the performance of alternative statistical methods for ranking medical providers and applies them to an analysis of data on satisfaction with mental health care from Group Health Cooperative, an integrated health care plan in Washington state. We explored performance of alternative statistical methods, varying patient volume, between-provider variation in patient characteristics, and between-provider variability in performance to reflect characteristics of providers and patients seen in a variety of health care settings. In this chapter, we provide a brief overview of the practice of provider profiling and an introduction to the statistical methods used for this purpose.

1.1 Overview of provider profiling

Provider profiling is a method of describing performance or patterns of health care provided by a single physician or physician group; the resulting profile is then compared to other norms based on practice, i.e., other physicians' profiles or to standards of practice. The basic concept of comparative profiling of medical providers dates back to Florence Nightingale (b. 1820), who pioneered the systematic collection, analysis, and dissemination of comparative hospital outcomes data in order to understand and improve performance (Smith, 2002). There has been a proliferation of studies conducted for this purpose since the 1960s.

Assessing the comparative performance of health care providers is now 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. Absolute performance standards, e.g. a national guideline, are normally unavailable (Normand et al., 1997), but 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 (Kak et al., 2001; Epstein et al., 2004). Moreover, policymakers may use performance data to assess whether core health care needs are being met (Hauck et al., 2003).

A potential approach to increasing the effectiveness of health care that has garnered interest among managed care organizations, purchasers, and researchers is the use of clinician financial incentives. Financial incentives may influence clinician performance in the form of either a higher payment for providing a particular service or a bonus for meeting a specified target. Recently, Meredith et al. (2011) used stakeholder input to select incentives for mental health clinicians and to conduct a feasibility test of an incentive-based program in a managed behavioral healthcare organization (MBHO). Program feasibility was assessed with case review and clinician surveys from a large independent practice association that contracted with the MBHO. They found that providing incentives for mental health clinicians was feasible and that the incentive program did increase clinician awareness.

The "pay for performance" (P4P) model is the application of the provider incentives idea on a broader scale, and provider profiling is increasingly used to identify providers exceeding performance thresholds, an integral part of the P4P concept. P4P is an emerging movement in health insurance that began in Britain and the United States. In the P4P model, physicians or hospitals receive more money if their quality measures exceed fixed thresholds or if their measures improve from year to year; this is a fundamental change from previous fee for service payments, which were instead dependent on the quantity of care. According to Epstein et al. (2004), a core goal of the P4P movement was to create financial incentives large enough to change the status quo rather than to simply reward "good" physicians and punish "bad" ones. By 2003, P4P was gaining widespread acceptance in health care.

Uses of provider profiling such as P4P differ from earlier uses, 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, little research on statistical methods for provider profiling has focused on classification accuracy.

1.2 Estimating provider performance

The basic goal of profiling methods is to estimate and compare provider performance using one or more proxy measures of performance. A provider's intrinsic ability is multifaceted and difficult to gauge, but excellent care tends to elicit certain responses in those receiving it, e.g. a feeling of satisfaction, a sense of rapport with the provider, and/or a willingness to return for future services. Each of these responses may be quantified on a numeric scale and combined across a provider's patients to provide an overall performance score. In this section we will discuss alternative approaches to generating performance scores and their pros and cons.

A simple method for assessing provider performance is to base estimation on the provider-specific sample average of a quality measure. Eisenstein et al. (2005) examined the use of average costs for creating physicians' economic profiles. For binary outcomes, this method corresponds to the proportion of patients with a positive response. When the measure of health status is mortality, a crude death rate can be obtained by summing the number of deaths per year per 1,000 people. Crude vital statistics were first calculated from the London Bills of Mortality to monitor plague deaths from the 17th century to the 1830s, and use of the crude death rate was discussed by Walker and Grusin (1959). Mortality rates are now widely used as a way to measure the quality of coronary artery bypass graft (CABG) surgery performed at a particular hospital.

However, there may be case-mix variation—differences in the population of patients treated by each provider—that makes such a comparison unfair. For example, some providers may see more difficult-to-treat patients by chance or even because they are known to be skilled providers. Also, certain subpopulations may be more predisposed to give positive feedback than others. Thus, ignoring patient characteristics related to

the patients' response to treatment, e.g. disease severity or gender, would put certain providers at a disadvantage in provider profiling if unadjusted measures are used.

Of the commonly used statistical methods for estimating provider performance, many strive for fairness by accounting for case-mix variation through a practice known as risk adjustment (Iezzoni, 1994; Rosenbaum, 1995). Confounding by case-mix variation occurs when a provider's performance rating is higher or lower than those of other providers due to reasons other than his/her performance. Two current caveats of P4P programs are that they could encourage physicians and hospitals to avoid high-risk patients in order to keep their performance scores up (McMahon et al., 2007) and could also put institutions and physicians attending to more vulnerable populations in resource-poor environments at a disadvantage (Casalino et al., 2007). As disparities in patient conditions are inevitable, profiling methods used in P4P or in any incentive-based program must take them into consideration so as to not encourage preferential patient admission.

Consequently, data quality and availability of information on confounders are key to obtaining unbiased estimates of provider performance. Normand and Shahian (2007) noted that in any profiling initiative, data quality is more important than choice of statistical models. Krumholz et al. (2006) recommended the use of a prospectively maintained clinical database containing core clinical variables. Indeed, if a key correlate of the outcome is unmeasured, then it would be impossible to obtain risk-adjusted results. Rosenbaum and Rubin (1983) demonstrated that conclusions about the association between treatment assignment and a binary outcome in an observational study, adjusted for categorical covariates, are sensitive to the omission of a binary confounder. Despite the importance of adjusting for confounders, it is usually impossible to adjust for them all. Even when data are relatively complete, other statistical concerns have been expressed regarding risk adjusted models, including inaccuracy of estimates from low-volume providers, multiple comparisons, and nonindependence of patients among providers (Normand and Shahian, 2007).

1.2.1 Indirect standardization

Indirect standardization is one means of addressing case-mix variation. This type of standardization produces a weighted sum of stratum-specific averages, with weights chosen such that the influence of case-mix effects can be reduced. It was first introduced as a method to assess geographical variations in adjusted death rates (Neison, 1844). Under this framework, 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 of a provider's patients is then calculated and the ratio of the observed to the expected score computed. Shahian and Normand (2008) used this method and data from the Society of Thoracic Surgeons National Database to compare hospital-specific, risk-standardized, 30-day all-cause mortality after isolated CABG surgery.

Indirect standardization is often used when the sample size is too small for direct standardization or when stratum specific rates are unavailable in the study population. The direct method, which utilizes weights taken from an external standard population, yields greater comparability across populations but requires more data (Daly and Bourke, 2000). The indirect method avoids the problem of imprecise estimates of stratum-specific rates by using stratum-specific rates from a standard population of sufficient size to derive expected counts in the study population. However, it may not be appropriate when samples sizes are small in strata given substantial weight, since weighting will magnify the instability of these stratum-specific estimates. Another limitation is that the choice of a standard population is not always obvious, and different choices may yield different results.

1.2.2 Hierarchical models

Hierarchical models (Bryk and Raudenbush, 1992) are widely used to evaluate providers using estimates of case-mix-adjusted performance (Glance et al., 2006; D'Errigo et al., 2007; Shahian et al., 2007). Under this framework, fixed effects are estimated for case-mix characteristics that may be associated with the outcome, and residual between-

provider variability is incorporated via a provider-level random effect.

Central to hierarchical modeling is the idea of shrinkage (James and Stein, 1961), which entails borrowing information across all units in the estimation of individual effects. Efron and Morris (1975) famously demonstrated that predictions for individual baseball players are much more accurate if the performances of all other players were taken into consideration, i.e., when each player's batting average was shrunk toward the group average.

Suppose we have data from k clusters of size n_i each and wish to estimate the mean of each cluster, μ_i. A no pooling approach would be to use cluster means; these maximum likelihood estimators are unbiased for μ_i but can have large standard errors if within-cluster variation is high. Complete pooling, i.e., calculating the grand mean, results in increased efficiency but also potentially large bias for μ_i if between-cluster variation is high. Stein (1955) proved the existence of a biased estimator that has lower mean squared error than the no pooling estimator. The random effects estimator, an extension of the James-Stein estimator (James and Stein, 1961), dominates the no pooling estimator in terms of mean squared error. For estimating each μ_i, it lies optimally between the cluster mean and the grand mean—close to the cluster mean if the clusters are very different and shrunk to the grand mean if the clusters are similar. Furthermore, estimates obtained from smaller clusters are shrunk toward the grand mean more than those from larger clusters, because for smaller clusters, there is weaker evidence that their means differ from the grand mean.

Hierarchical models have many advantages in provider profiling. They can incorporate both patient-level and provider-level characteristics and account for within-provider correlation in patient outcomes. They not only adjust for variation in case-mix, but they 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).

Furthermore, in hierarchical modeling, variation in sample size is addressed by borrowing information across providers to stabilize estimates for providers who have relatively few subjects. The random effects distribution determines the degree to which

information is borrowed; smaller variance implies a greater degree of sharing.

However, hierarchical models are sensitive to distributional assumptions. Random effects (provider effects in the profiling context) are assumed to be drawn from an appropriate distribution—usually a normal distribution (Normand et al., 1997; DeLong et al., 1997; Austin, 2002) with mean 0. The normality assumption can lead to severe shrinkage of estimates for providers with few patients, so methods that use flexible distributional assumptions (e.g. Paddock et al. (2006)) are more suitable if estimates are to be used for detecting outliers. Moreover, hierarchical modeling assumes "exchangeability" (Bernardo and Smith, 1994) of the providers, meaning that provider performance arises from a common distribution for all providers. This notion is unjustified if performance depends on patient or provider characteristics, although if these characteristics are included in the regression model, then only conditional exchangeability is required.

1.2.3 Bayesian approaches to profiling

Bayesian hierarchical modeling is increasingly being advocated for use in profiling medical care (Berlowitz et al., 2002). In Bayesian inference, prior distributions are specified for unknown parameters of a chosen model to capture one's beliefs before seeing the data, and Bayes' theorem is applied to obtain posterior distributions for the parameters of interest (Savage, 1954). Thomas et al. (1994) used empirical Bayes for estimating hospital-specific mortality rates. Their fully model-based formulation was found to produce accurate estimates and resolved the problem of multiple comparisons.

Browne and Draper (2006) examined Markov chain Monte Carlo (MCMC) estimation in hierarchical modeling via a large simulation study of the properties of quasi-likelihood and Bayesian estimation methods in the random effects logistic regression model. They noted that estimates achieved using both proper and improper priors on the variance parameters were close to unbiased and had coverage probabilities closer to nominal at all levels when compared to the two quasi-likelihood methods under study.

Caveats exist concerning more recent methodologies. Berlowitz et al. (2002) examined the use of Bayesian hierarchical modeling in profiling nursing homes on the rate of pressure ulcer development. Though they cited several advantages of Bayesian hierarchical modeling over standard statistical techniques in terms of estimating risk-adjusted ulcer rates and identifying outliers, they warned that appropriate inferences from Bayesian models require careful model-checking to ensure that the model is not misspecified. Furthermore, sophisticated methods that appeal to statisticians may not yield results that are transparent to all stakeholders in provider profiling. The Bayesian semi-parametric models developed in Ohlssen et al. (2007) to detect "true" outlying health care providers were critiqued by Racz and Sedransk (2010) as being innovative but having "complicated and cumbersome" parts. The framework proposed by Ohlssen and colleagues is based on three stages of analysis—an exploratory stage that involves using a normal random-effects model and approximate cross-validation p-values to identify potential outlying providers, a modeling stage that involves performing detailed sensitivity analysis of potential outliers, and a confirmatory stage that involves applying the Bonferroni correction to each p-value. While these are all statistically sound approaches, they are not easily utilized or understood by a non-statistical audience. Transparency is truly an important consideration due to the widespread implications and varied consumers of provider profiling.

1.3 Previous evaluations of methods for provider profiling

Previous research has compared methods for estimating provider performance. Austin et al. (2003) compared the performance of indirect standardization and hierarchical models for identifying hospitals with outlying (higher than acceptable or lower than acceptable) mortality rates with respect to sensitivity, specificity, and positive predictive value. They found that when the distribution of hospital-specific log-odds of death was normal, random-effects models had greater specificity and positive predictive value than fixed-effects models. However, fixed-effects models had greater sensitivity than random-effects models. Racz and Sedransk (2010) compared Bayesian and likelihood-

based frequentist approaches to identifying outlying hospitals using coronary artery bypass graft surgery data released by the New York State Department of Health. The six methods under comparison were indirect standardization, fixed effects models, and random effects models, each estimated via Bayesian and frequentist methods. They found that for medium and large hospital volumes, indirect standardization and the random effects model yielded almost identical results. This is because, as noted in Normand et al. (1997), the estimators for hospitals with large patient volume are only slightly shrunk toward the group mean, even when they are quite different from it. For low volume hospitals, however, there was a marked reduction in outlier detection using the random effects model as compared to using indirect standardization in simulations, as estimators for these hospitals experienced greater degrees of shrinkage than those for larger hospitals.

Previous research has also examined the statistical uncertainty in provider profiling. Ohlssen et al. (2007) described a hierarchical modeling framework for identifying unusual performance in health care providers and observed that using rank-based criteria not only has the problem of great uncertainty in estimates of provider ranks but also is conceptually problematic even when estimates are reasonably precise. Their paper tried to distinguish between an estimation approach and a hypothesis testing approach for describing unusual performance. The former is based on a single encompassing model for identifying extreme cases, while the latter uses a fairly simple model to describe most cases, with interest in assessing whether the few are truly unusual. They concluded that ranking might indicate which providers are at the extremes but cannot address whether the worst provider is actually worse than would be expected by chance.

Goldstein and Spiegelhalter (1996) noted that ranks are particularly sensitive to sampling variability, so there is a need to quantify their uncertainty when they are used to make comparisons between institutions, with an easily interpretable example being a graph of ranks and their intervals. To illustrate this problem, the authors considered examination scores taken from a single-year cohort of students nested within schools. The following two-level variance components model was fit.

$$y_{ij} = \beta_0 + u_j + e_{ij},$$
$$\mathrm{var}(u_j) = \sigma_u^2,$$
$$\mathrm{var}(e_{ij}) = \sigma_e^2,$$

where y_{ij} is the examination score for the ith student in the jth school, u_j is the residual for the jth school, and e_{ij} is the residual for the ith student in the jth school. The residuals are assumed to be mutually independent with zero means. This model yields posterior estimates $\hat{u}_j$ and $\mathrm{var}(\hat{u}_j)$, which were then used for comparisons between institutions. After applying this model to real data and plotting the school intercept residual estimates in ascending order with their 95% intervals, they commented that in terms of pairwise comparisons, the majority of the schools could not be distinguished from one another due to overlapping intervals. An adjusted model showed that an estimated two-thirds of all possible comparisons among 325 schools do not allow separation. The authors concluded that such rankings may allow some institutions at the extremes to be isolated for further study but should not be used to make definitive judgments on individual institutions.

Goldstein's view on the importance of quantifying statistical uncertainty in provider profiling was echoed in Davidson et al. (2007), who considered estimating the amount of uncertainty in measuring relative quality to be a policy relevant issue, as low levels of confidence with which hospitals are assigned to the top percentile of ranks means that such top performers may not actually be worthy of rewards and/or public recognition. Their main metric for portraying the uncertainty in estimates of the true relative performance of hospitals was the 95% credible interval about the mean rank derived from the Bayesian models for each medical condition. They concluded that identifying relative quality from simple ranks based on annual composite scores will impact smaller institutions to a greater extent than larger institutions. The dramatic inverse relationship found between hospital size and ranking uncertainty suggested that in profiling studies, institutions with smaller patient volume will be wrongly penalized or credited more frequently than will larger institutions.

Paddock and Louis (2011) further highlighted the risk of misclassifying providers as exceptionally good or poor performers when uncertainty in statistical benchmark estimates is ignored. Statistical benchmarks are performance standards derived from existing data for the purpose of setting performance targets (e.g. the top 90th percentile of provider performance). The typical benchmarking approach is to use one data set to derive the statistical benchmark estimate and then to use a second data set to obtain provider-specific performance estimates and to compare them with the previously determined benchmark. The authors' goal was to identify extreme (e.g. top 10%) provider performance; they developed alternative empirical distribution function (EDF) estimates for univariate provider-specific parameters based on order statistics of MCMC samples drawn from the posterior distribution of provider-specific parameters, and their approach produced the optimal estimates with respect to integrated squared error loss. Their rationale for considering the EDF was that while approximate uncertainty bounds for commonly used statistical benchmarks can be obtained by using the posterior mean and variance, it is more desirable to avoid relying on the Bayesian central limit theorem.

1.4 Our work

Our research was motivated by the use of profiling for identifying top-performing providers in the setting of mental health care. Group Health, an integrated health care system in Washington state and North Idaho, has been developing incentives for its contracted network providers and plans to fully implement their use for quality improvement this year (GHC, 2012). However, a study by Katon et al. (2000) had found no important differences in quality of care or patient outcomes, when hierarchical logistic regression models were used to estimate mental health care provider quality. Motivated by the need for accurately identifying top and bottom providers, meanwhile entertaining the possibility that little variation exists in the quality of care delivered by health care providers (Katon et al., 2000; Krein et al., 2002), we explored the performance of alternative statistical methods for identifying providers exceeding

a percentile-based threshold. We evaluated both classification accuracy and uncertainty, as recent literature has suggested that an understanding of uncertainty is key in provider profiling studies (Goldstein and Spiegelhalter, 1996; Davidson et al., 2007; Paddock and Louis, 2011).

In selecting methods for comparison, we took into consideration ease of interpretation, since straightforward methods are more widely used and are more readily accepted by non-statisticians. For the same reason, we chose to investigate classification of providers based on a single outcome measure. The statistical issues highlighted in previous research led us to scrutinize primarily how the methods compare in realistic scenarios involving low patient volumes and high case-mix variation and additionally in more ideal, albeit unrealistic, situations. Holding other factors constant, we expected all methods to discriminate more successfully when greater differences exist in provider performance, and we wanted to quantify the gain, if any, of methods that account for clustering and case-mix variation over methods that do not. As low patient volume, presence of case-mix variation, and low between-provider variability are common in the medical field, we are particularly interested in how methods compare in such close-to-life settings.

In Chapter 2, we first describe the three classification methods under comparison and then compare their classification accuracy for estimating provider performance using a simulation study. 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 also quantified the uncertainty of the estimated provider ranks using a Bayesian estimation method. In Chapter 3, we apply these statistical methods to data from Group Health, using the results of simulation studies to guide our interpretation of results, and compare the performance of alternative methods. In Chapter 4, we summarize our findings and discuss future directions.

Chapter 2

METHODS EVALUATION

This chapter introduces three methods for provider profiling and the three measures of classification performance upon which they were compared. Here we also describe the design of our simulation study and present results for an investigation of the relationship between classification performance and mean patient volume, degree of between-provider variability, and strength of case-mix effects.

2.1 Notation and definitions

Suppose there are N providers, with the ith provider having n_i patients and the jth patient of the ith provider having outcome measure, Y_{ij}. We focus on the case of binary outcome measures. However, similar considerations apply to continuous measures. Let $\boldsymbol{X}_{ij}$ represent a vector of patient characteristics, such as age, gender, and disease severity. Let θ_i be the ith provider's true performance score, and assume

$$\theta_1, \ldots, \theta_N | (\mu, \sigma^2) \sim D(\mu, \sigma^2),$$

where D is assumed to be a known distribution. In numerical examples below, θ_i is assumed to follow a normal distribution. In principle, $D(,)$ need not be normal. While a provider's true performance is independent of patient characteristics, these characteristics may affect patient outcomes and consequently impact observed performance.

We assume, without loss of generality, that in a given sample of providers, the provider with the poorest performance will have the smallest θ value. The true rank of the ith provider, based on his/her performance, θ_i, is given by

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

We estimate the rank of the ith provider by

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

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

2.2 Methods for estimating provider performance

A straightforward method of assessing provider performance is to use the provider-specific sample average as the estimate. In the case of binary outcomes, this corresponds to the proportion of subjects with a positive response. Formally, we define

$$\hat{\theta}_i^{(E)} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}.$$

Below, we refer to this as the empirical method. This method is appealing for its simplicity and transparency. However, no adjustment is made for differences in case-mix between providers or possible instability of estimated ranks for providers with low patient volumes.

Indirect standardization is a simple alternative which allows for adjustment on the basis of measured patient characteristics. This method provides an estimate of provider performance standardized by an expected performance measure computed based on patient characteristics. This standardized performance score is computed as the ratio of the sum of observed outcomes to the sum of expected outcomes had these patients been treated by an average provider in the reference population (Shahian and Normand, 2008). 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:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta},$$

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

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

where $\hat{p}_{ij} = \exp(\boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}})/(1+\exp(\boldsymbol{X}_{ij}^T\hat{\boldsymbol{\beta}}))$ is the fitted probability for the jth patient of the ith provider from the 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 last method that we consider is a hierarchical logistic regression model assuming common effects of patient characteristics across providers and allowing for between-provider variability in the outcome via a random effect. We use a random intercepts model of the form

$$\log\left(\frac{p_{ij}^*}{1-p_{ij}^*}\right) = \alpha_i + \boldsymbol{X}_{ij}^T\boldsymbol{\beta},$$

where Y_{ij} is now assumed Bernoulli distributed with mean p_{ij}^*, which is assumed to depend on both patient characteristics and provider. The provider-specific random intercepts, α_i, are assumed to arise from a common distribution

$$\alpha_i \sim \mathcal{N}(\mu, \sigma^2).$$

μ and σ^2 denote the mean and variance, respectively, of the distribution of random intercepts. In this method, provider performance is assessed using the estimated provider-specific random effects

$$\hat{\theta}_i^{(R)} = \hat{\alpha}_i,$$

with higher values corresponding to better performance. The hierarchical model can also be used to estimate the degree of variability between providers, $\hat{\sigma}$.

2.3 Estimator precision

The performance of rankings depends on the precision of our estimates for θ_i. In the case of the empirical method, $\hat{\theta}_i^{(E)}$, the variance is given by $\theta_i(1-\theta_i)/n_i$. Note that $\theta_i(1-\theta_i)$ can be thought of as the degree of within provider variability. For the indirect standardization method, variance of the estimator is $n_i\theta_i(1-\theta_i)/(\sum_j \hat{p}_{ij})^2$, where $\hat{p}_{ij}$ is the estimated probability of a positive response for the jth observation for provider i, assumed to be fixed and known.

In the case of the random effects approach, an explicit formula for the estimator is not available. However, we can use the inverse Fisher information as a measure of the asymptotic variance. Specifically, let $g(\theta_i)$ represent a function mapping provider performance, θ_i, to the probability of a positive response for an individual subject. For instance, in the case of the logistic random intercepts model we assume a logistic relationship between θ_i and p_{ij}. For ease of presentation, we assume that all subjects for provider i have the same risk factor values. However, similar results follow allowing for differences in risk factors within providers. We further assume that θ_i is normally distributed with mean μ and variance σ^2. Under this formulation the likelihood for θ_i is given by

$$L \propto \exp\left(-\frac{1}{2\sigma^2}(\theta_i - \mu)^2\right) g(\theta_i)^{\sum_j Y_{ij}} (1 - g(\theta_i))^{n_i - \sum_j Y_{ij}}.$$

From this likelihood we can derive the score function

$$\frac{\partial \ell}{\partial \theta_i} = -\frac{1}{\sigma^2}(\theta_i - \mu) + \frac{\partial g}{\partial \theta_i}\left(\frac{\sum_j Y_{ij} - n_i g(\theta_i)}{g(\theta_i)(1 - g(\theta_i))}\right).$$

Differentiating the score function again with respect to θ_i, we see that the negative inverse Fisher information is

$$-\left(\frac{\partial^2 \ell}{\partial \theta_i^2}\right)^{-1} = \frac{\sigma^2}{\sigma^2 h\left(\sum_j Y_{ij}; \theta_i, n_i\right) + 1}, \tag{2.1}$$

where

$$h\left(\sum_j Y_{ij}; \theta_i, n_i\right) = \frac{\partial}{\partial \theta_i}\left(\frac{\partial g}{\partial \theta_i}\left(\frac{\sum_j Y_{ij} - n_i g(\theta_i)}{g(\theta_i)(1 - g(\theta_i))}\right)\right).$$

From Equation (2.1) we can see that the variance of $\hat{\theta}_i^{(R)}$ will be directly proportional to σ^2 and inversely proportional to n_i.

The variance expressions for the three methods for estimating provider performance indicate that increasing the number of evaluations available for an individual provider, n_i, will improve the precision of the estimator. However, in the case of $\hat{\theta}_i^{(I)}$ and $\hat{\theta}_i^{(R)}$ the improvement in precision associated with increasing sample size is attenuated by other factors. Specifically, in the case of $\hat{\theta}_i^{(I)}$ improvements in precision are proportional to $(\sum_j \hat{p}_{ij})^2/n_i$. Thus for providers with low expected performance scores,

sample size will improve performance more slowly. In the case of $\hat{\theta}_i^{(R)}$, when there is a high degree of between provider variability (σ^2), performance will improve as a function of n_i. When between provider variability is low, σ^2 will be the dominant factor affecting estimator performance.

The precision of each of these estimators of provider performance must be evaluated with respect to the total amount of between provider variability. If σ^2 is small then even modest levels of uncertainty in our estimators may make it impossible to correctly rank providers. In contrast, when σ^2 is high we need not estimate θ_i with great precision in order to place providers in the correct rank order. Based on the forms of the variance for the three methods of estimating provider performance, we anticipate that the performance of the methods for ranking providers will depend on sample size and the true degree of between-provider variability.

The uncertainty of the estimated provider ranks can be quantified using a Bayesian estimation method. For our analysis of mental health satisfaction survey data, diffuse prior distributions were assumed for the parameters in each model. Specifically, for classification via the empirical method, each provider's mean was assumed to be normally distributed, centered at 0 with variance 1,000. All other fixed effects in logistic regression models were assumed to be normally distributed, with mean 0 and variance 1,000. 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, a single chain was used for the Gibbs sampler, and the burn-in phase consisted of 1,000 iterations. After burn-in, one thousand iterations were run for the empirical method; four thousand were run for the other two methods, thinned by a factor of 4. A 95% credible interval for the ranks was constructed based on this sample. WinBUGS 1.4.3 (Lunn et al., 2000) was used for Bayesian estimation.

Below we explore the impact of sample size and between-provider variability in simulation studies using values representative of those expected to be encountered

in studies of health care provider performance. In Chapter 3, we use the previously described methods to rank Group Health providers and quantify the uncertainty of their ranks.

2.4 Assessing performance of classification

As mentioned in Chapter 1, 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 their ranks. However, in general, true performance is unknown, and an estimate such as those previously described must be used to carry out classification. Suppose we are interested in identifying $100p_0\%$ of N providers as "top" performers. We assign each provider a rank based on an estimate of their performance. Top performers are defined as those with ranks $N(1-p_0)+1$ to N. The classification performance of an estimator can be assessed in a variety of ways. In this study, we focus on three measures of classification performance:

1. Sensitivity, defined as the proportion of top performing providers that are correctly classified as such, i.e.,

$$\frac{1}{Np_0}\sum_{i=1}^{N} I_{[R(\hat{\theta}_i)>N(1-p_0),\, R(\theta_i)>N(1-p_0)]}.$$

2. Specificity, defined as the proportion of non-top performers that are correctly classified, i.e.,

$$\frac{1}{N(1-p_0)}\sum_{i=1}^{N} I_{[R(\hat{\theta}_i)\leq N(1-p_0),\, R(\theta_i)\leq N(1-p_0)]}.$$

3. Root mean squared error (RMSE) of the provider ranks, defined as

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}[R(\hat{\theta}_i)-R(\theta_i)]^2}.$$

2.4.1 Simulation study design

Because providers' true performance measures are unobservable, classification performance can only be assessed using Monte Carlo methods. We simulated random data representing N providers, with Poisson-distributed patient volumes. Values for N, patient volume, and between-provider variability were motivated by data on satisfaction with mental health care from Group Health (see Section 3.1). We generated each provider's true performance, α_i, from a $\mathcal{N}(0, \sigma^2)$ distribution for a range of values of σ^2. 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 value ranging from -5 to 5 for the jth patient of the ith provider from a truncated $\mathcal{N}(m_i, 1)$ distribution, with m_i, the ith provider's mean value, simulated from a truncated $\mathcal{N}(0, 2)$ 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

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

where α_i is the true performance score for the ith provider, X_{ij} is the covariate value for the jth patient of the ith provider, and γ is the odds ratio associated with a one unit difference in X_{ij}. For each simulated data set, we estimated performance using each method, classified providers into the top 20% based on each performance estimate, and then evaluated classification performance using the measures discussed in Section 2.4. All results are based on 1,000 replications, which yielded a maximum Monte Carlo standard error of 0.004 for sensitivity and specificity and 0.05 for RMSE. Finally, we simulated a data set of comparable size and variability as survey data from Group Health mental health care providers (described in Section 3.1) to summarize both performance and agreement among the three methods. We used R 2.10.1 (R Development Core Team, Vienna, Austria) for our simulations and data analyses.

2.5 Simulation study results

We investigated the relationship between classification performance and mean patient volume, strength of case-mix effects, and degree of between-provider variability for a simulated sample of 50 providers with patient volumes of 20, 30, 40, 50, 70, and 100. In the absence of case-mix variation, i.e., $\gamma = 1$, all three methods performed similarly in terms of sensitivity and specificity at each mean patient volume considered (Figure 2.1). Furthermore, all methods showed improvements in sensitivity and specificity as patient volume increased at each level of between-provider variability. When case-mix effects were present, little improvement was evident with increased patient volume for the empirical method at all levels of between-provider variability (Figures 2.2 and 2.3).

We report results for trends in sensitivity, specificity, and RMSE as functions of the effect of case-mix variation and degree of between provider variability for simulated data for 50 providers with a mean of 50 patients per provider (Table 2.2). We found similar trends for more extreme patient volumes of 20 and 100 (Table 2.1 and Table 2.3). For $\gamma = 1$, the three classification methods were comparable in terms of sensitivity, specificity and RMSE. Sensitivity for the empirical method decreased dramatically as the effect of case-mix variation increased. Between the two methods that adjusted for patient differences, classification based on random effects was robust to the presence of increased effects of case-mix. For instance, when $\sigma^2 = 1$, increasing γ from 1 to 2 resulted in a decrease in sensitivity for the indirect standardization method of 10%, while sensitivity of the random effects method decreased by only 3%. RMSE was lower for the random effects method compared to the empirical and indirect standardization methods when the effect of case-mix variability was greater than 1.

Overall, classification performance and ranking precision improved in terms of sensitivity, specificity, and RMSE as between-provider variability increased. Under very low levels of between-provider variability ($\sigma^2 = 0.01$), all three methods performed very poorly, with sensitivity below 40%. Sensitivity and specificity improved as between-provider variability increased, with the random-effects method slightly outperforming other methods in terms of sensitivity and specificity for nearly all simulated scenar-

ios. We found that sensitivity was quite poor for even moderate levels of between-provider variability. Precision of the provider rankings, as indicated by RMSE, decreased substantially as between-provider variability increased. Differences between the random effects method and indirect standardization at larger cluster sizes were more pronounced for $\gamma = 2$ due to the plateauing of the latter method's sensitivity and specificity.

Finally, we evaluated agreement of ranks from the three methods using simulated data for 55 providers with mean patient volume of 30 per provider, low between-provider variability ($\sigma^2 = 0.04$), and no case-mix effects ($\gamma = 1$). These parameters were motivated by Group Health satisfaction survey data, which are described in detail in Chapter 3.

Agreement of methods was strong between ranking by indirect standardization and by the other two methods and less so between the random effects and empirical methods. Despite the strength of agreement, Figure 2.4 reveals substantial misclassification of true top and bottom 20% providers. This is consistent with the low sensitivity and specificity associated with the case of moderate between-provider variability and no case-mix variation in Table 2.2.

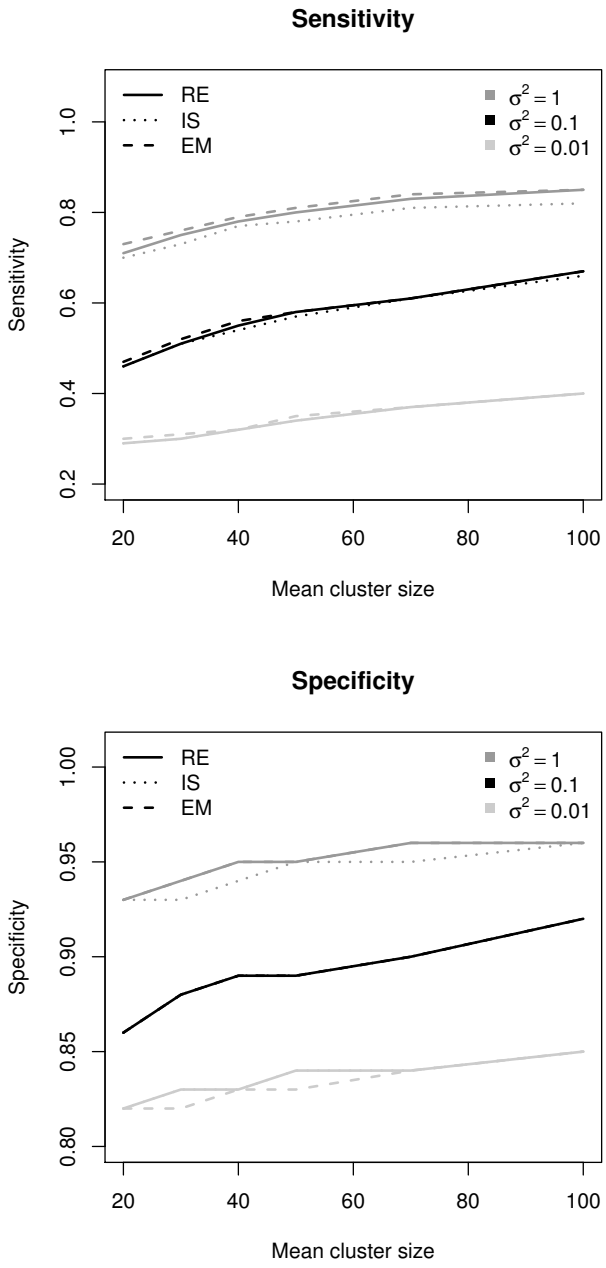

Figure 2.1: **Sensitivity and specificity for no case-mix effect.** Estimated sensitivity and specificity of classification into top 20% of providers for three methods, empirical (EM), indirect standardization (IS), and random effects (RE), for 50 providers, with values of mean patient volume varying from 20 to 100, effect of patient variation, γ, equal to 1, and degree of between-provider variability, σ^2, equal to 0.01, 0.1, and 1.

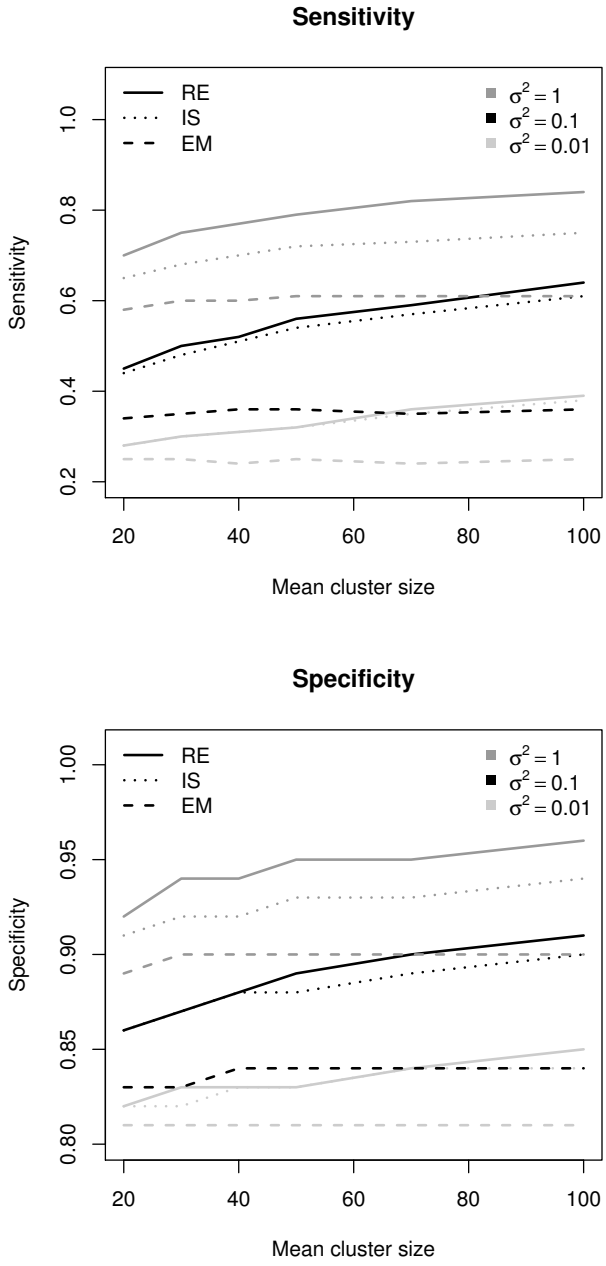

Figure 2.2: **Sensitivity and specificity for moderate case-mix effect.** Estimated sensitivity and specificity of classification into top 20% of providers for three methods, empirical (EM), indirect standardization (IS), and random effects (RE), for 50 providers, with values of mean patient volume varying from 20 to 100, effect of patient variation, γ, equal to 1.5, and degree of between-provider variability, σ^2, equal to 0.01, 0.1, and 1.

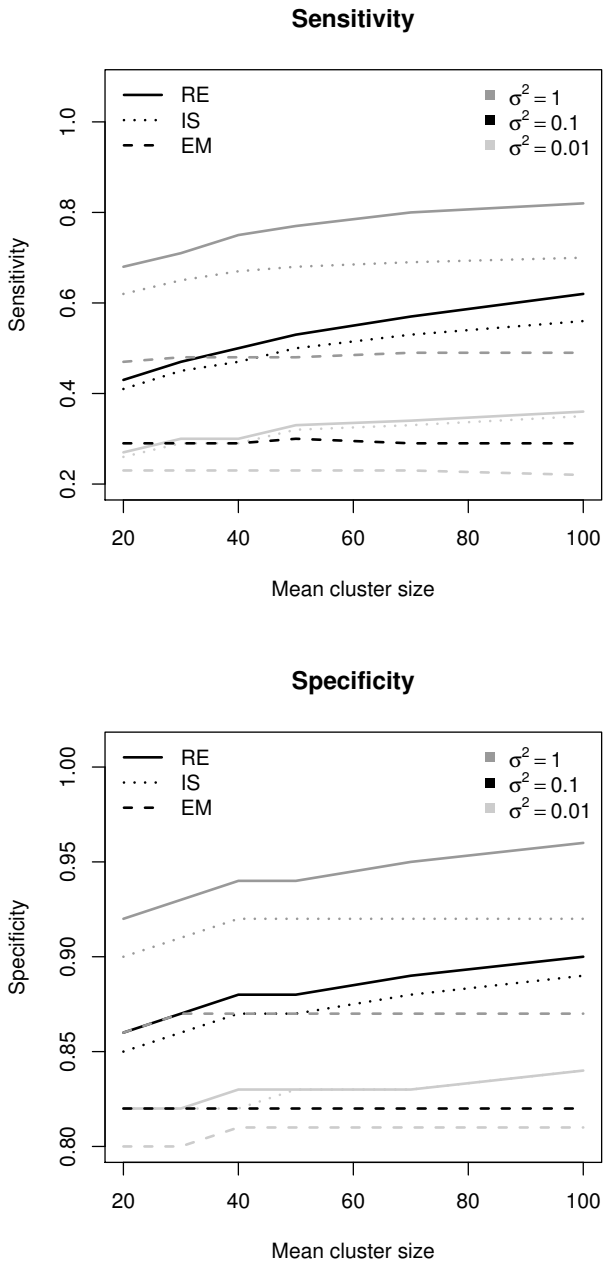

Figure 2.3: **Sensitivity and specificity for high case-mix effect.** Estimated sensitivity and specificity of classification into top 20% of providers for three methods, empirical (EM), indirect standardization (IS), and random effects (RE), for 50 providers, with values of mean patient volume varying from 20 to 100, effect of patient variation, γ, equal to 2, and degree of between-provider variability, σ^2, equal to 0.01, 0.1, and 1.

Table 2.1: **Classification performance for mean patient volume of 20.** Estimated sensitivity and specificity for classification into top 20% of providers and RMSE of ranks based on three classification methods, empirical (E), indirect standardization (I), and random effects (R), for varying strength of effect of patient variation (γ), three levels of between-provider variability (σ^2), and mean patient volume of 20 per provider.

	$\gamma = 1$			$\gamma = 1.5$			$\gamma = 2$		
	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1
	Sensitivity								
$\hat{\boldsymbol{\theta}}^{(E)}$	0.29	0.46	0.71	0.28	0.45	0.70	0.27	0.43	0.68
$\hat{\boldsymbol{\theta}}^{(I)}$	0.30	0.47	0.73	0.25	0.34	0.58	0.23	0.29	0.47
$\hat{\boldsymbol{\theta}}^{(R)}$	0.29	0.46	0.70	0.28	0.44	0.65	0.26	0.41	0.62
	Specificity								
$\hat{\boldsymbol{\theta}}^{(E)}$	0.82	0.86	0.93	0.82	0.86	0.92	0.82	0.86	0.92
$\hat{\boldsymbol{\theta}}^{(I)}$	0.82	0.86	0.93	0.81	0.83	0.89	0.80	0.82	0.86
$\hat{\boldsymbol{\theta}}^{(R)}$	0.82	0.86	0.93	0.82	0.86	0.91	0.82	0.85	0.90
	RMSE								
$\hat{\boldsymbol{\theta}}^{(E)}$	18.22	13.74	7.02	18.29	14.13	7.42	18.57	14.64	8.08
$\hat{\boldsymbol{\theta}}^{(I)}$	18.20	13.71	6.90	19.24	16.88	10.90	19.68	18.20	13.79
$\hat{\boldsymbol{\theta}}^{(R)}$	18.23	13.78	7.11	18.32	14.29	8.00	18.66	14.97	9.01

Table 2.2: **Classification performance for mean patient volume of 50.** Estimated sensitivity and specificity for classification into top 20% of providers and RMSE of ranks based on three classification methods, empirical (E), indirect standardization (I), and random effects (R), for varying strength of effect of patient variation (γ), three levels of between-provider variability (σ^2), and mean patient volume of 50 per provider.

	$\gamma = 1$			$\gamma = 1.5$			$\gamma = 2$		
	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1
	Sensitivity								
$\hat{\boldsymbol{\theta}}^{(E)}$	0.35	0.58	0.81	0.25	0.35	0.61	0.23	0.30	0.48
$\hat{\boldsymbol{\theta}}^{(I)}$	0.34	0.57	0.78	0.33	0.53	0.72	0.31	0.49	0.68
$\hat{\boldsymbol{\theta}}^{(R)}$	0.34	0.57	0.80	0.33	0.55	0.80	0.32	0.53	0.77
	Specificity								
$\hat{\boldsymbol{\theta}}^{(E)}$	0.83	0.89	0.95	0.81	0.84	0.90	0.81	0.82	0.87
$\hat{\boldsymbol{\theta}}^{(I)}$	0.83	0.89	0.95	0.83	0.88	0.93	0.83	0.87	0.92
$\hat{\boldsymbol{\theta}}^{(R)}$	0.83	0.89	0.95	0.83	0.89	0.95	0.83	0.88	0.94
	RMSE								
$\hat{\boldsymbol{\theta}}^{(E)}$	16.91	10.95	4.74	19.15	16.61	10.19	19.60	18.09	13.45
$\hat{\boldsymbol{\theta}}^{(I)}$	16.95	11.05	5.06	17.21	11.72	6.16	17.56	12.61	7.19
$\hat{\boldsymbol{\theta}}^{(R)}$	16.93	11.04	4.86	17.15	11.46	5.18	17.43	12.14	5.73

Table 2.3: **Classification performance for mean patient volume of 100.** Estimated sensitivity and specificity for classification into top 20% of providers and RMSE of ranks based on three classification methods, empirical (E), indirect standardization (I), and random effects (R), for varying strength of effect of patient variation (γ), three levels of between-provider variability (σ^2), and mean patient volume of 100 per provider.

	$\gamma = 1$			$\gamma = 1.5$			$\gamma = 2$		
	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1	σ^2 = 0.01	σ^2 = 0.1	σ^2 = 1
	Sensitivity								
$\hat{\boldsymbol{\theta}}^{(E)}$	0.40	0.67	0.85	0.39	0.64	0.84	0.36	0.62	0.82
$\hat{\boldsymbol{\theta}}^{(I)}$	0.40	0.67	0.85	0.25	0.36	0.61	0.22	0.29	0.49
$\hat{\boldsymbol{\theta}}^{(R)}$	0.40	0.66	0.82	0.38	0.61	0.75	0.35	0.56	0.70
	Specificity								
$\hat{\boldsymbol{\theta}}^{(E)}$	0.85	0.92	0.96	0.85	0.91	0.96	0.84	0.90	0.96
$\hat{\boldsymbol{\theta}}^{(I)}$	0.85	0.92	0.96	0.81	0.84	0.90	0.81	0.82	0.87
$\hat{\boldsymbol{\theta}}^{(R)}$	0.85	0.92	0.96	0.84	0.90	0.94	0.84	0.89	0.92
	RMSE								
$\hat{\boldsymbol{\theta}}^{(E)}$	15.48	8.82	3.68	15.86	9.27	3.88	16.26	10.03	4.38
$\hat{\boldsymbol{\theta}}^{(I)}$	15.44	8.73	3.57	19.11	16.44	9.98	19.66	18.02	13.25
$\hat{\boldsymbol{\theta}}^{(R)}$	15.50	8.85	3.96	15.96	9.68	5.27	16.46	10.69	6.41

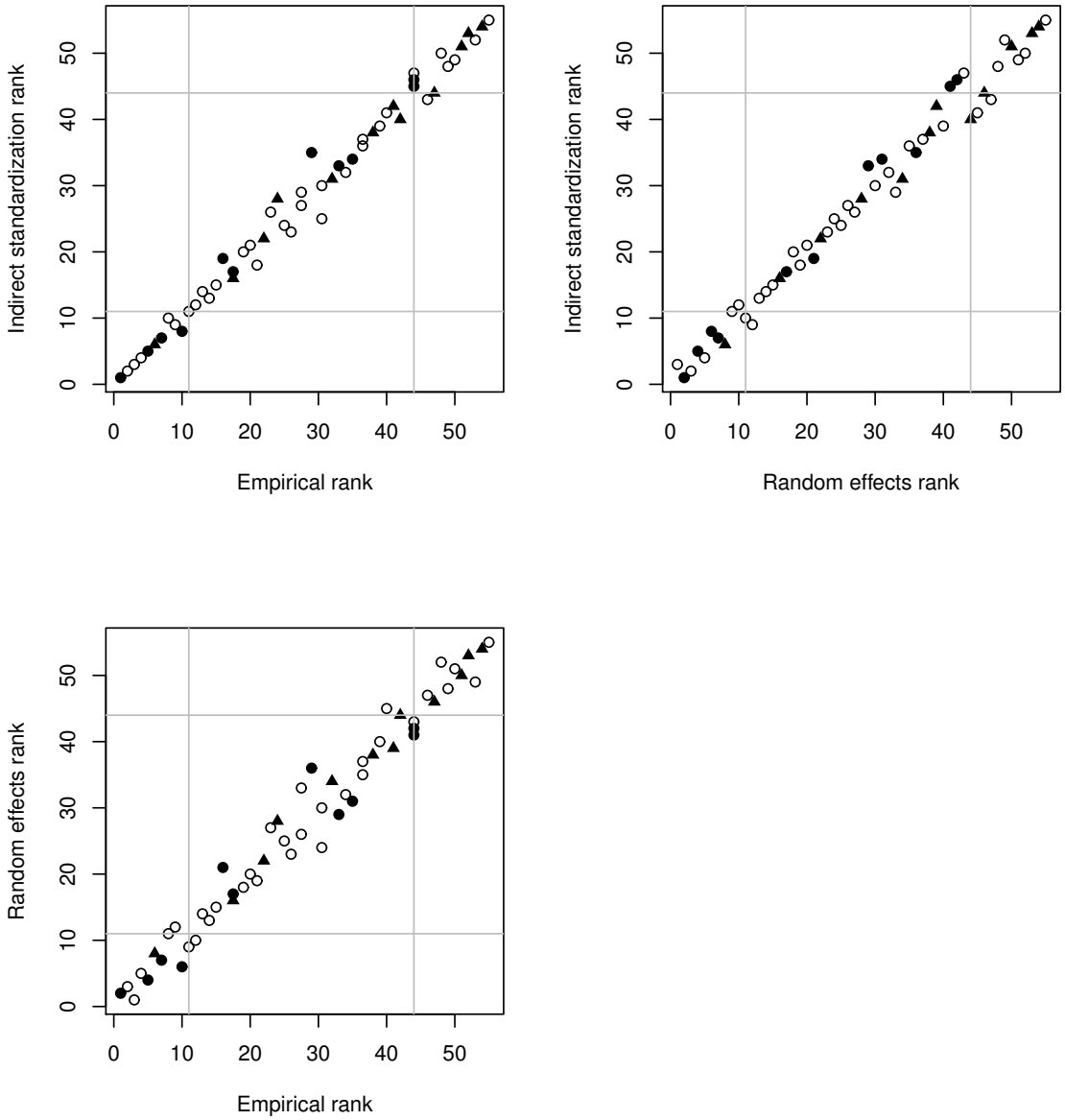

Figure 2.4: **Agreement of rankings in a simulated setting.** Agreement of rankings of 55 providers based on three methods for estimating provider performance in a simulated setting ($\gamma = 1$, $\sigma^2 = 0.04$, mean patient volume of 30 per provider). Triangles indicate true top 20%, empty circles indicate true middle 60%, and filled circles indicate true bottom 20%. Horizontal and vertical lines represent cutpoints for the 20th and 80th percentiles of classification ranks.

Chapter 3

APPLICATION TO GROUP HEALTH DATA

In this chapter, we first describe the mental health provider satisfaction survey data from Group Health and then use the methods described in the previous chapter to estimate provider performance and rank Group Health providers. Uncertainty of the estimated ranks was quantified using a Bayesian approach. Our analysis indicated that in the event of low between-provider variability among Group Health providers, classification based on the three methods was relatively consistent but that its performance would have been shown to be poor if a gold standard method were available, as demonstrated by our simulation study in Chapter 2.

3.1 Description of Group Health provider satisfaction data

We compared the performance of our 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 500,000 members in Washington state and northern Idaho. Within these states, the Group Health enrollment is similar to the population of the serviced area in terms of income, educational attainment, and representation of different racial and ethnic groups. 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 the cities of 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 March, 2008 and February, 2010 from pa-

tients 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 as the basis for determining which providers receive bonuses or are recommended for additional training (Crosier et al., 2012). Patients rated this on a 5-point scale, ranging from 1, poor, to 5, excellent. 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.

These data have been previously described by Simon et al. (2009). Although the response rate was only 34% , previous analyses have found that demographic characteristics and visit patterns of respondents and non-respondents were similar (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 in Section 2.1 to estimate provider performance and rank providers from Group Health. We also quantified the uncertainty of the estimated provider ranks using a Bayesian estimation method.

3.2 Group Health data analysis results

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. The number of surveys per provider ranged from 9 to 69, with a median of 30 for group providers and ranged from 5 to 33 for network providers, with a median of 10. Table 3.1 summarizes the patient characteristics of those who submitted these surveys. Patients who rated group and network providers were comparable in age, sex, diagnosis, and

Table 3.1: **Description of Group Health enrollees.** Description of Group Health enrollees who submitted satisfaction ratings for group or network providers.

Patient characteristics	Group providers (N=1,742)	Network providers (N=1,522)
Age (mean, sd)	48.7 (15.1)	47.9 (14.1)
Male (n, %)	483 (27.7%)	375 (24.6%)
Depression diagnosis	955 (54.8%)	857 (56.3%)
Anxiety diagnosis	411 (23.6%)	397 (26.1%)
Bipolar/Psychosis diagnosis	85 (4.9%)	110 (7.3%)
Other diagnosis	291 (16.7%)	158 (10.3%)
Had prior visit(s)	613 (35.2%)	249 (16.4%)
$\geq$24 mos. enrollment	1,297 (74.5%)	1,056 (69.4%)

enrollment duration, though network providers were rated more frequently by new patients. Comparison between the two groups may be biased by selection of different patients into care with group and network providers (Simon and Ludman, 2010).

Using unadjusted random effects models we estimated the standard deviation of the provider random effects in each sample to quantify between-provider variability. We anticipated that there would be less variability between group providers, who practice in a common setting and participate in common training, than among network providers. The standard deviation of the random effects for group model providers was 0.202 and for network providers was 0.552. Based on random effects from this model, the estimated proportion of a provider's patients rating their satisfaction as "excellent" ranged from 56.5% to 67.3% for group providers and from 41.0% to 81.1% for network providers. Figure 3.1 shows the distribution of these estimates for group and network providers. The difference between the top 20% and bottom 20% among group providers was only 3.5%, while among network providers top and bottom performers differed by 12.3%.

We estimated provider performance for the two samples using each of the three

methods described above. Simon and Ludman (2010) noted that dropout from psychotherapy was also typically associated with younger age, minority race or ethnicity, no previous mental health treatment, and a longer time between screening and the initial visit. Therefore, in the indirect standardization and random effects methods we adjusted for patient age in years, a continuous variable, to demonstrate the degree of discrepancy between an unadjusted method (the empirical method) and methods that adjusted for variability in providers' patient populations. In the group sample we found that age was not significantly associated with satisfaction (odds ratio = 1.001, p = 0.18). We therefore expected that in this context the effect of adjusting for age on classification should be minimal. However, age was significantly associated with satisfaction in the network sample (odds ratio = 1.002, p = 0.04).

Because no gold standard classification is available, we compared provider rankings based on the three methods. Figures 3.2 and 3.3 show ranks for each provider based on performance scores estimated using each of the three methods. For both group and network providers, ranks based on the empirical and indirect standardization methods agree closely with almost no disagreement in classification into the top and bottom 20% of providers between the two methods. Agreement between the random effects method and the other two methods was poorer. Agreement between methods appears similar for the group and network samples, despite the greater degree of variability among the network providers and the presence of a significant effect of a patient characteristic, age, on the outcome.

We plotted 95% credible intervals for the estimated ranks for group providers (Figure 3.4). 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 empirical or indirect standardization approaches. Similar results were found for network providers (Figure 3.5).

In this chapter, we compared classification of providers into the top 20% based on the three methods and found high levels of agreement between methods. As our simulation study has shown, consistent results do not necessarily imply accurate classi-

fication for any of the three methods when between-provider variability is low. We also found substantial uncertainty in the estimated ranks of both group and network providers.

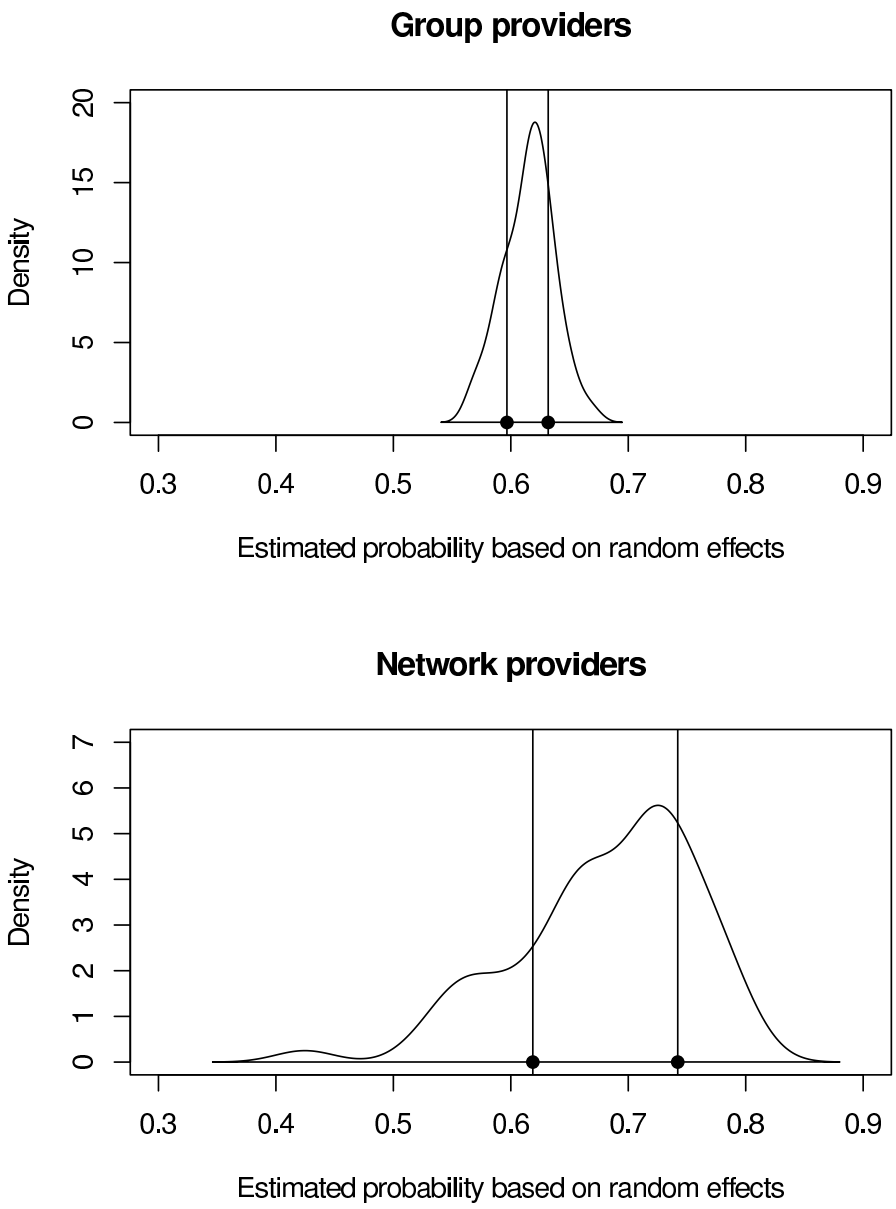

Figure 3.1: **Distribution of proportion of "excellent" ratings.** Estimated distribution of proportion of "excellent" ratings for 55 group providers and 128 network providers. Vertical lines represent cutpoints for the 20th and 80th percentile of providers.

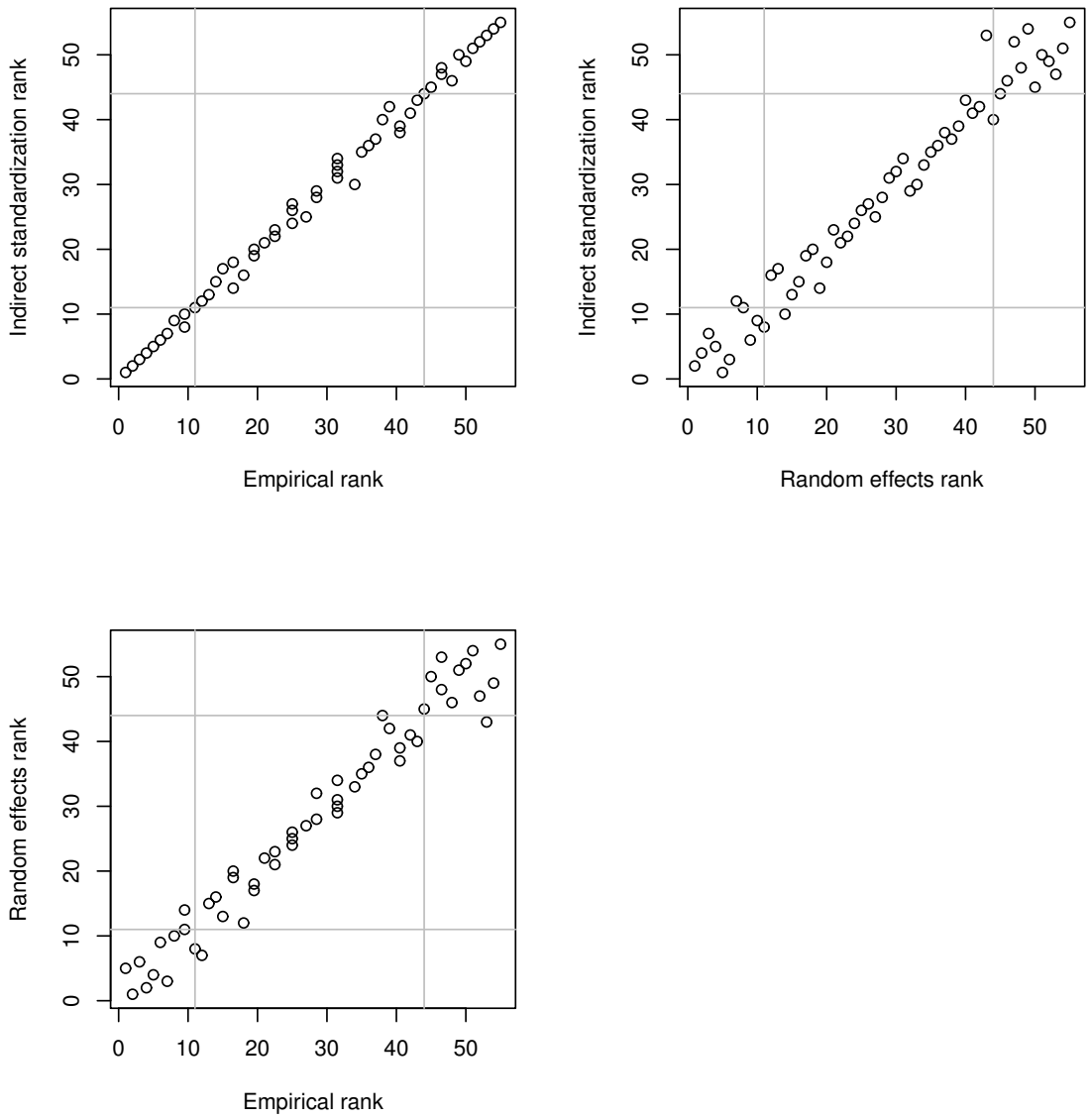

Figure 3.2: **Agreement of rankings of group model providers.** Agreement of rankings of 55 group model providers based on three methods for estimating provider performance. Horizontal and vertical lines represent cutpoints for the 20th and 80th percentile of providers.

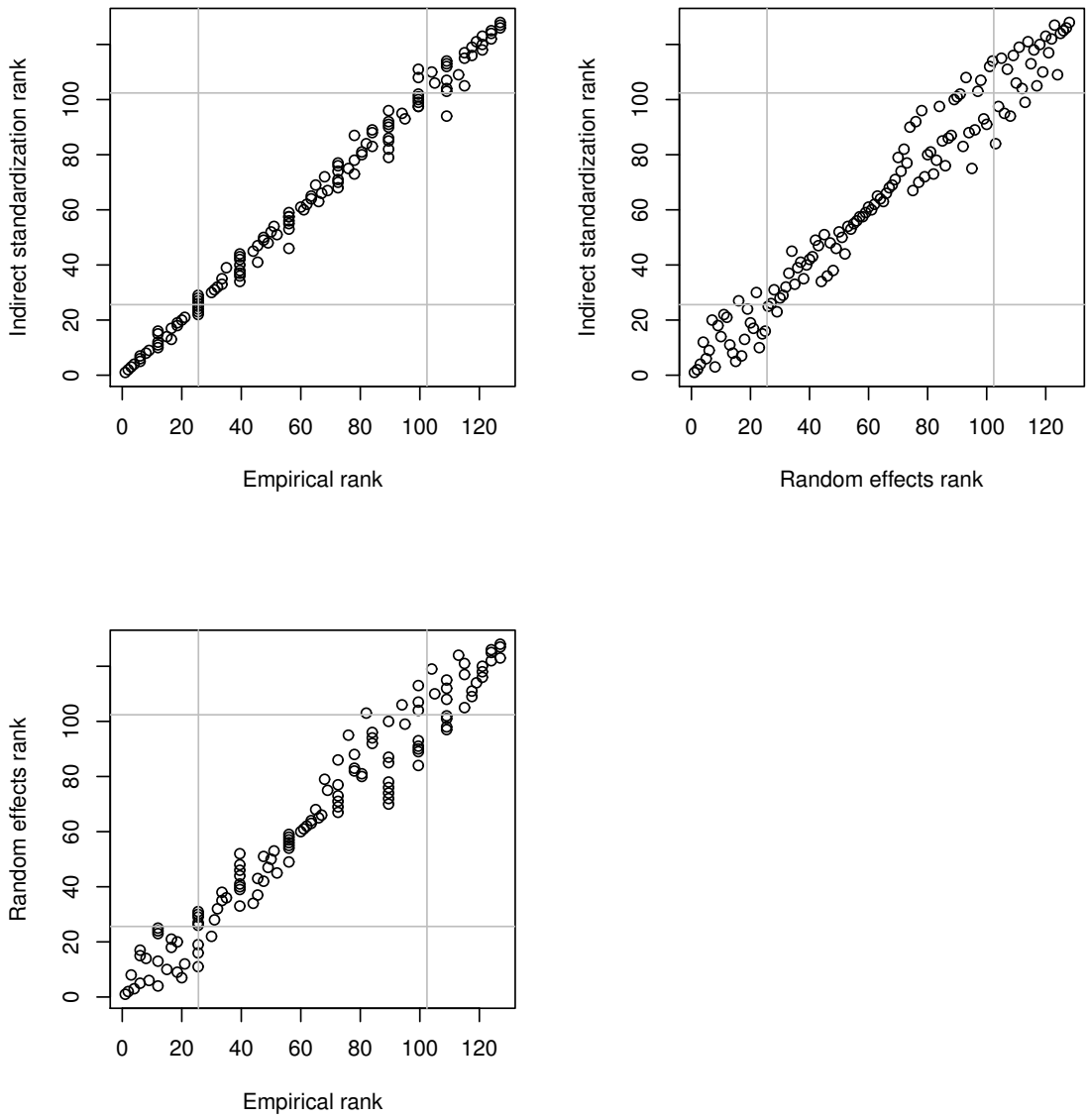

Figure 3.3: **Agreement of rankings of network model providers.** Agreement of rankings of 128 network model providers based on three methods for estimating provider performance. Horizontal and vertical lines represent cutpoints for the 20th and 80th percentile of providers.

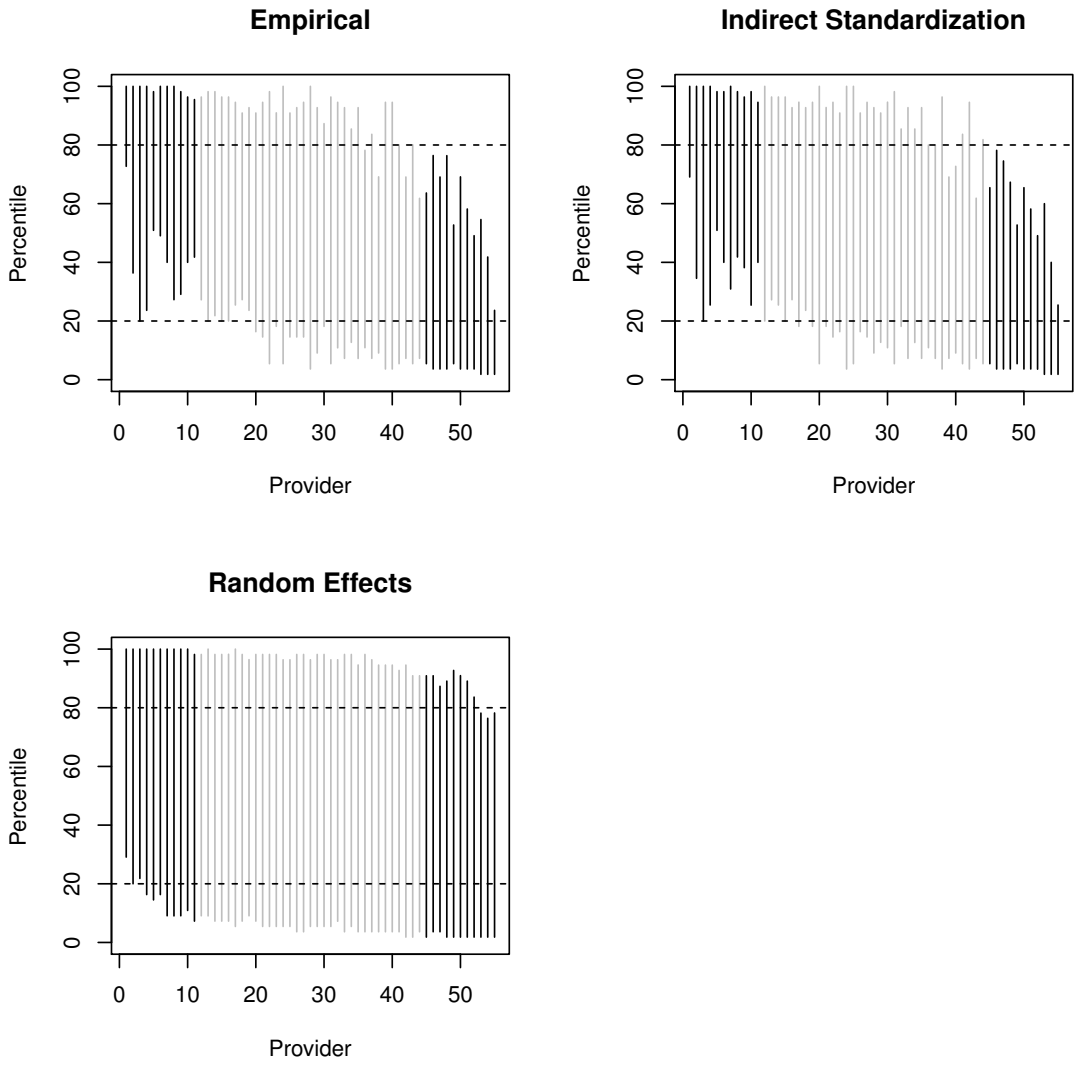

Figure 3.4: **Credible intervals for ranks of group model providers.** 95% credible intervals for ranks of 55 group model providers based on three methods for estimating provider performance. Providers are indexed in decreasing order of estimated performance, as determined by the posterior rank. Intervals for the estimated top and bottom 20% of providers are displayed in black. Horizontal lines represent 80th and 20th percentile of providers.

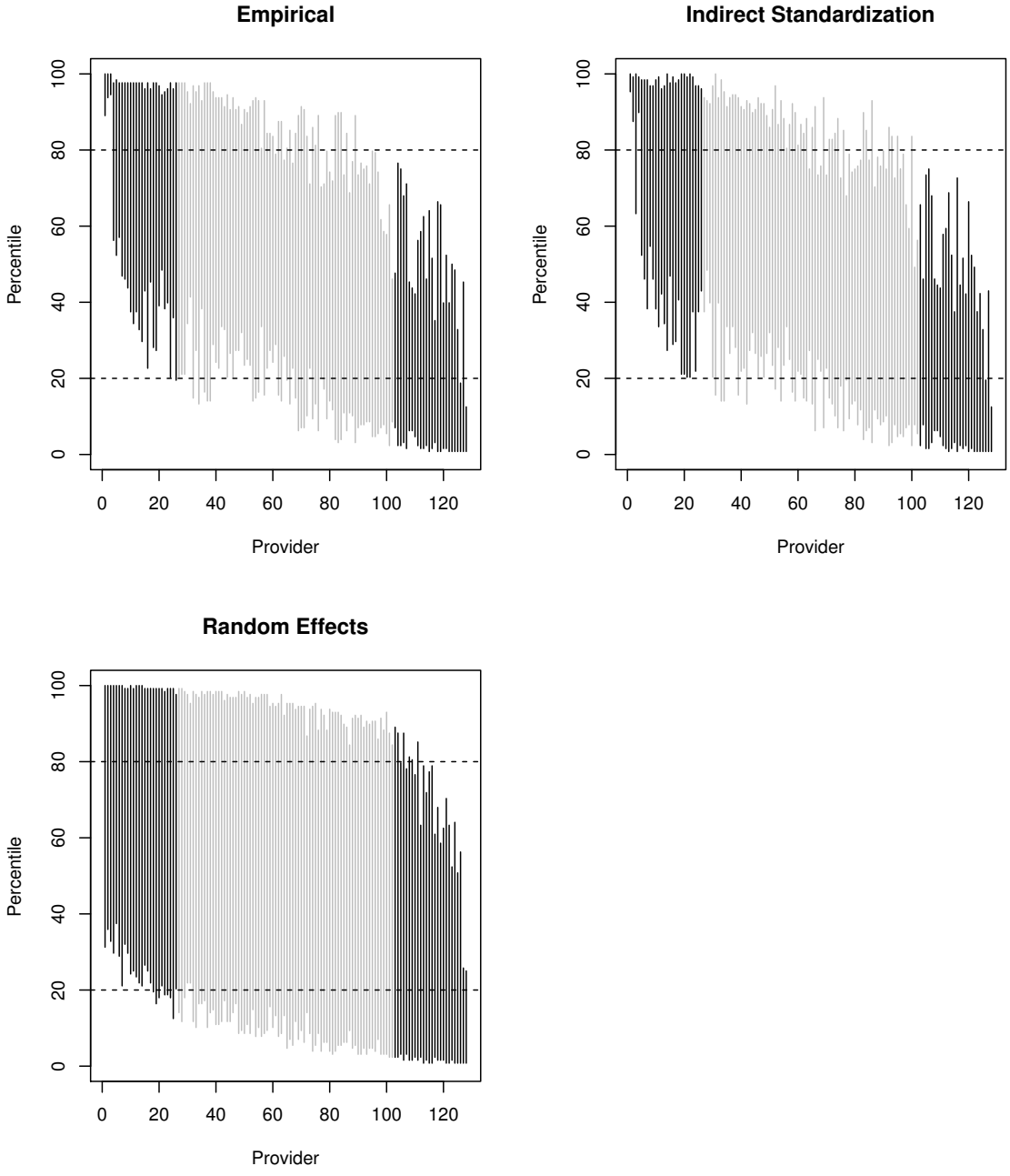

Figure 3.5: **Credible intervals for ranks of network model providers.** 95% credible intervals for ranks of 128 network model providers based on three methods for estimating provider performance. Providers are indexed in decreasing order of estimated performance, as determined by the posterior rank. Intervals for the estimated top and bottom 20% of providers are displayed in black. Horizontal lines represent 80th and 20th percentile of providers.

Chapter 4

DISCUSSION

We compared classification accuracy, sensitivity and specificity, as well as precision of three methods for estimating provider performance—an empirical method, an indirect standardization method, and a hierarchical model. This chapter is a discussion of findings from our simulations and Group Health data analysis and concludes with some remarks on the future direction of our research.

4.1 Discussion of simulations

Our simulation study indicated that all three methods perform poorly when little between-provider variability exists. Our simulations covered a range of plausible values for patient volume, number of providers, variation in patients seen among providers, and between-provider variability in the setting of mental health provider performance. In this context, past studies have shown that minority race or ethnicity and younger age are significantly negatively associated with patient satisfaction (Simon and Ludman, 2010). Failing to adjust for such covariates could result in misleading inference. Our simulation studies demonstrated that the empirical method is inappropriate when the distribution of patient characteristics differs across providers and affects patient ratings of providers. While we only included a single covariate capturing differences in patient characteristics, our results can be generalized to data with multiple patient-level characteristics that are associated with a binary outcome. Small sample size, large variations in patient characteristics across providers, and low between-provider variability all contribute to poor classification accuracy. In the presence of substantial between-provider variability, all three methods will perform adequately even if sample sizes for individual providers are relatively small or variability in patient populations exists between providers. However, when between-provider

variability is low none of these methods perform well, regardless of large patient volumes or homogeneity of patient characteristics across providers. Thus, low between-provider variability is the greatest obstacle to successful classification via the methods we explored.

When between-provider variability is low, classification by any of the three methods investigated here is unreliable. Indeed, the degree of variability among providers should be considered before classification is undertaken, because in these low variability cases, performance is so similar across providers that classification into top and bottom performing groups may not be meaningful (Lockwood et al., 2002). For moderate patient volumes and levels of variability, hierarchical models or indirect standardization may be undertaken. Statistical methods that appropriately account for differences in patient characteristics that are associated with the outcome are critical to successful estimation of provider performance, yet even these methods cannot yield satisfactory results when providers are similar. We conclude that exploration of between provider variability is an important first step in any provider profiling analysis.

4.2 Discussion of Group Health data analysis

Our analysis of data on satisfaction with mental health care indicated that there is relatively little variability between providers both within the Group Health practice and the contracted community network. While the network providers had greater between-provider variability than group providers (either because network providers are inherently more heterogeneous in their practice or because it was driven by greater variability in their patient volume), it was not sufficient for top performing providers to be accurately identified based on comparison of performance in simulations with this degree of variability. We compared classification based on the three methods and found high levels of agreement between methods. Based on our simulation study, we anticipate that the three methods would provide relatively consistent but erroneous classifications in this setting. We caution against making inference on top- and bottom-performers in the setting of low between-provider variability. Agreement in classifica-

tion across methods should not be interpreted as an indication of accurate classification. Our simulation studies showed that in this scenario all methods are likely to agree on the same erroneous classifications.

Consistent with previous literature on the behavior of ranks in profiling studies (Goldstein and Spiegelhalter, 1996; Davidson et al., 2007), estimates of variability in Group Health 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 this high degree of uncertainty in this setting, it is not surprising that misclassification is common.

Consequently, any time a profiling initiative is undertaken, we advise quantifying the uncertainty of the procedure. Standard errors, confidence intervals, or credible intervals of estimated ranks can be used as a means of assessing whether or not classification is reliable in a given setting, with one example being the 95%credible interval about the mean rank derived from Bayesian models (Davidson et al., 2007).

Previous research on performance of mental health providers has suggested that hierarchical models can be used to identify top/bottom performing providers using both cross-sectional (Landrum et al., 2003) and longitudinal (Bronskill et al., 2002) data. Our analysis of data from Group Health indicated that in our sample, between provider variability was insufficient to allow for reliable classification of providers. Performance of estimation methods may be improved if repeated measures for individual subjects were available (Okiishi et al., 2003, 2006; Wampold and Brown, 2005). This would allow for more sensitive adjustment for random variation attributable to patients than is possible with single measures per subject and is likely to perform better than estimates based on single observations per subject. However, repeated assessments are often unavailable.

4.3 Future direction and final comments

There are a number of additional research directions for studies of statistical methods for provider profiling. We assumed that random effects were normally distributed and only simulated normally distributed random effects. We could consider alternative data generating models in our future simulations, e.g. allowing the provider random effects to arise from a non-normal or a normal mixture distribution. This would allow us to see whether the random effects model will achieve the highest sensitivity and specificity even when the random effects distribution is misspecified.

Misspecification of the random effects distribution has been found to cause estimation problems for the hierarchical modeling approach. Heagerty and Kurland (2001) computed the asymptotic bias in the maximum likelihood estimators for the parameters in a logistic mixed model in four instances of random-effect model misspecification, one of which—relevant to our simulation study design—was the use of maximum likelihood to fit a simple Gaussian random intercepts model when random intercepts are in fact gamma-distributed. They found that the asymptotic relative bias of the intercept estimate was on the order of 30% for highly skewed distributions and even greater when between-cluster heterogeneity was substantial. MuCulloch 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, although this is not expected to greatly impact provider rankings based on estimated provider effects.

We illustrated classification of providers using various methods based on percentiles. This approach identifies a prespecified number, the top 20% of all providers considered, as top-performers. As critiqued by Berlowitz et al. (2002), this approach may be problematic in that there may not be any evidence that a provider with outlying performance is different from the norm; this was also addressed in Ohlssen et al. (2007). Satisfaction survey responses are typically skewed to the positive end (Lebow, 1982),

so we have reason to believe that mental health providers, usually holding at least a Master's degree and sufficiently trained, tend to deliver care of comparable quality. However, our current classification scheme treats a provider at the 79th percentile very differently from one at the 80th.

Because Bayesian hierarchical modeling provides not only a point estimate of performance but also the probability that a provider's true performance exceeds a fixed threshold, we could alternatively classify providers based on whether they exceed some posterior probability threshold of lying in the extremes. The group of providers for which this probability is high would then be the ones to identify as top-performers. This reward system is less quota-based and more merit-based, although the threshold is still an arbitrarily assigned value rather than a separation of outliers and the norm.

Several authors have published on the use of posterior tail probabilities in profiling. Austin and Brunner (2008) used Monte Carlo methods to assess the accuracy of posterior tail probabilities derived from Bayesian hierarchical regression models for identifying hospitals with higher than acceptable mortality, and they demonstrated that the use of posterior tail probabilities was the Bayes' rule associated with generalized 1-0 loss functions. Austin (2008) additionally developed Bayes' rules for squared error loss and absolute error loss when Bayesian hierarchical regression models are used to identify hospitals with unacceptably high mortality. Austin (2008) also investigated the impact of assuming each of these three loss functions on the number of hospitals identified as having unacceptably high mortality and found it to be minimal. We therefore need not consider multiple loss functions in our future work.

A variety of methods are currently used in profiling studies, and we saw that the sophisticated random effects model—despite its theoretical and practical advantages—performed no better than unadjusted means in realistic settings. Since low between-provider variability and low patient volume are prevalent in health care settings, one must maintain a certain degree of skepticism. As Goldstein and Spiegelhalter (Goldstein and Spiegelhalter, 1996) so aptly said, profiling results should be treated as suggestive rather than definitive.

BIBLIOGRAPHY

Austin, P. C. (2002). A comparison of Bayesian methods for profiling hospital performance. *Medical Decision Making* **22,** 163–172.

Austin, P. C. (2008). Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals. *BMC Medical Research Methodology* **8,** 1–11.

Austin, P. C., Alter, D. A., and Tu, J. V. (2003). The use of fixed- and random-effects models for classifying hospitals as mortality outliers: A Monte Carlo assessment. *Medical Decision Making* **23,** 526–539.

Austin, P. C. and Brunner, L. J. (2008). Optimal Bayesian probability levels for hospital report cards. *Health Services and Outcomes Research Methodology* **8,** 80–97.

Berlowitz, D. R., Christiansen, C. L., Brandeis, G. H., Ash, A. S., Kader, B., Morris, J., and Moskowitz, M. (2002). Profiling nursing homes using Bayesian hierarchical modeling. *Journal of the American Geriatrics Society* **50,** 1126–1130.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester.

Bronskill, S. E., Normand, S.-L. T., Landrum, M. B., and Rosenheck, R. A. (2002). Longitudinal profiles of health care providers. *Statistics in Medicine* **21,** 1067–1088.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **1,** 473–513.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical linear models*. Sage, Newbury Park, CA.

Burgess, J. F., Christiansen, C. L., Michalak, S. E., and Morris, C. N. (2000). Medical profiling: improving standards and risk adjustments using hierarchical models. *Journal of Health Economics* **19,** 291–309.

Casalino, L. P., Elster, A., Eisenberg, A., Lewis, E., Montgomery, J., and Ramos, D. (2007). Will pay-for-performance and quality reporting affect health care disparities? *Health Affairs* **26,** 405–414.

Crosier, M., Scott, J., and Steinfeld, B. (2012). Improving satisfaction in patients receiving mental health care: a case study. *Journal of Behavioral Health Services & Research* **39,** 42–54.

Daly, L. E. and Bourke, G. J. (2000). *Interpretation and uses of medical statistics*. Blackwell Publishing, Oxford, UK.

Davidson, G., Moscovice, I., and Remus, D. (2007). Hospital size, uncertainty, and pay-for-performance. *Health Care Financing Review* **29,** 45–57.

DeLong, E. R., Peterson, E. D., DeLong, D. M., Muhlbaier, L. H., Hackett, S., and Mark, D. B. (1997). Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine* **16,** 2645–2664.

D'Errigo, P., Tosti, M. E., Fusco, D., Perucci, C. A., and Seccareccia, F. (2007). Use of hierarchical models to evaluate performance of cardiac surgery centres in the Italian CABG outcome study. *BMC Medical Research Methodology* **7,** 29–37.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70,** 311–319.

Eisenstein, E. L., Bethea, C. F., Muhlbaier, L. H., Davidian, M., Peterson, E. D., Stafford, J. A., and Mark, D. B. (2005). Surgeons' Economic Profiles: Can We Get the "Right" Answers? *Journal of Medical Systems* **29,** 111–124.

Epstein, A. M., Lee, T. H., and Hamel, M. B. (2004). Paying Physicians for High-Quality Care. *New England Journal of Medicine* **350,** 406–410.

GHC (2012). Quality plan and program description 2012 update. Website. https://www.ghc.org/about_gh/Quality/2012-quality-plan.pdf.

Glance, L. G., Dick, A., Osler, T. M., Li, Y., and Mukamel, D. B. (2006). Impact of Changing the Statistical Methodology on Hospital and Surgeon Ranking: The Case of the New York State Cardiac Surgery Report Card. *Medical Care* **44,** 311–319.

Goldstein, H. and Spiegelhalter, D. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society Series A-Statistics in Society* **159,** 385–409.

Hauck, K., Rice, N., and Smith, P. (2003). The influence of health care organisations on health system performance. *Journal of Health Services Research & Policy* **8,** 68–74.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88,** 973–985.

Iezzoni, L. I. (1994). *Risk Adjustment for Measuring Health Care Outcomes*. Health Administration Press, Ann Arbor, MI.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1,** 361–379.

Kak, N., Burkhalter, B., and Cooper, M.-A. (2001). Measuring the Competency of Healthcare Providers. *Quality Assurance Project* **2,** 1–28.

Katon, W., Rutter, C. M., Lin, E., Simon, G., Von Korff, M., Bush, T., Walker, E., and Ludman, E. (2000). Are There Detectable Differences in Quality of Care or Outcome of Depression across Primary Care Providers? *Medical Care* **38,** 552–561.

Krein, S. L., Hofer, T. P., Kerr, E. A., and Hayward, R. A. (2002). Whom Should We Profile? Examining Diabetes Care Practice Variation among Primary Care Providers, Provider Groups, and Health Care Facilities. *Health Services Research* **37,** 1159–1180.

Krumholz, H. M., Brindis, R. G., Brush, J. E., Cohen, D. J., Epstein, A. J., Furie, K., Howard, G., Peterson, E. D., Rathore, S. S., Smith, S. C., Spertus, J. A., Wang, Y., and Normand, S.-L. T. (2006). Standards for statistical models used for public reporting of health outcomes - An American Heart Association scientific statement from the quality of care and outcomes research interdisciplinary writing group - Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council - Endorsed by the American College of Cardiology Foundation. *Circulation* **113,** 456–462.

Landrum, M. B., Normand, S.-L. T., and Rosenheck, R. A. (2003). Selection of related multivariate means: Monitoring psychiatric care in the department of veterans affairs. *Journal of the American Statistical Association* **98,** 7–16.

Lebow, J. L. (1982). Consumer Satisfaction with Mental Health Treatment. *Psychological Bulletin* **91,** 244–259.

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

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS–a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10,** 325–337.

Marshall, M. N., Shekelle, P. G., Leatherman, S., and Brook, R. H. (2000). The Public Release of Performance Data: What Do We Expect to Gain? A Review of the Evidence. *Journal of the American Medical Association* **283,** 1866–1874.

McMahon, L. F., Hofer, T. P., and Hayward, R. A. (2007). Physician-level P4P-DOA? Can quality-based payment be resuscitated? *American Journal of Managed Care* **13,** 233–236.

Meredith, L. S., Branstrom, R. B., Azocar, F., Rikes, R., and Ettner, S. L. (2011). A Collaborative Approach to Identifying Effective Incentives for Mental Health Clinicians

to Improve Depression Care in a Large Managed Behavioral Healthcare Organization. *Administration and Policy in Mental Health* **38,** 193–202.

MuCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Statistical Science* **26,** 388–402.

Neison, F. G. P. (1844). On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts. *Journal of the Statistical Society of London* pages 40–68.

Normand, S.-L. T., Glickman, M. E., and Gatsonis, C. A. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association* **92,** 803–814.

Normand, S.-L. T. and Shahian, D. M. (2007). Statistical and clinical aspects of hospital outcomes profilling. *Statistical Science* **22,** 206–226.

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007). A hierarchical modelling framework for identifying unusual performance in health care providers. *Journal of the Royal Statistical Society Series A-Statistics in Society* **170,** 865–890.

Okiishi, J., Lambert, M. J., Eggett, D., Nielsen, L., and Dayton, D. D. (2006). An Analysis of Therapist Treatment Effects: Toward Providing Feedback to Individual Therapists on Their Clients' Psychotherapy Outcome. *Journal of Clinical Psychology* **9,** 1157–1172.

Okiishi, J., Lambert, M. J., Nielsen, S. L., and Ogles, B. M. (2003). Waiting for Supershrink: An Empirical Analysis of Therapist Effects. *Clinical Psychology and Psychotherapy* **10,** 361–373.

Paddock, S. M. and Louis, T. A. (2011). Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *Journal of the Royal Statistical Society Series C-Applied Statistics* **60,** 575–589.

Paddock, S. M., Ridgeway, G., Lin, R. H., and Louis, T. A. (2006). Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Computational Statistics & Data Analysis* **50,** 3243–3262.

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

Rosenbaum, P. (1995). *Observational Studies*. Springer-Verlag, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B-Methodological* **45,** 212–218.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York, second edition.

Shahian, D. M. and Normand, S.-L. T. (2008). Comparison of "Risk-Adjusted" Hospital Outcomes. *Circulation* **117,** 1955–1963.

Shahian, D. M., Silverstein, T., Lovett, A. F., Wolf, R. E., and Normand, S.-L. T. (2007). Comparison of clinical and administrative data sources for hospital coronary artery bypass graft surgery report cards. *Circulation* **115,** 1518–1527.

Simon, G., Rutter, C., Crosier, M., Scott, J., Operskalski, B. H., and Ludman, E. (2009). Are Comparisons of Consumer Satisfaction With Providers Biased by Nonresponse or Case-Mix Differences? *Psychiatric Services* **60,** 67–73.

Simon, G. E. and Ludman, E. J. (2010). Predictors of Early Dropout from Psychotherapy for Depression in Community Practice. *Psychiatric Services* **61,** 684–689.

Smith, P. C. (2002). Measuring health system performance. *European Journal of Health Economics* **60,** 145–148.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1,** 197–206.

Thomas, N., Longford, N. T., and Rolph, J. E. (1994). Empirical Bayes methods for estimating hospital-specific mortality-rates. *Statistics in Medicine* **13,** 889–903.

Walker, A. R. P. and Grusin, H. (1959). Coronary Heart Disease and Cerebral Vascular Disease in the South African Bantu – Examination and Discussion of Crude and Age-specific Death Rates. *American Journal of Clinical Nutrition* **7,** 264–270.

Wampold, B. E. and Brown, G. S. (2005). Estimating Variability in Outcomes Attributable to Therapists: A Naturalistic Study of Outcomes in Managed Care. *Journal of Consulting and Clinical Psychology* **5,** 914–923.