Medical profiling: improving standards and risk adjustments using hierarchical models

James F. Burgess Jr. a, *, Cindy L. Christiansen b, Sarah E. Michalak c, Carl N. Morris c

a Management Science Group, Department of Veterans Affairs (518 / MSG), 200 Springs Road, Bedford, MA 01730, USA
b Harvard Medical School and Harvard Pilgrim Health Care, USA
c Statistics Department at Harvard University, Cambridge, MA, USA

Abstract

The conclusions from a profile analysis to identify performance extremes can be affected substantially by the standards and statistical methods used and by the adequacy of risk adjustment. Medically meaningful standards are proposed to replace common statistical standards. Hierarchical regression methods can handle several levels of random variation, make risk adjustments for the providers’ case-mix differences, and address the proposed standards. These methods determine probabilities needed to make meaningful profiles of medical units based on standards set by all appropriate parties. Published by Elsevier Science B.V.

JEL classification: I18; C11; L15
Keywords: Profiling standards; Hierarchical models; Regression-to-the-mean; Risk adjustment

1. Introduction

Measuring and understanding differences in health care provider performance are drawing increasing attention from government agencies providing care or
subsidies to purchase care, from firms providing health care benefits to employees, from managed care organizations and other insurers, and from individual consumers selecting health care providers. Health providers and insurance companies, in particular, have increased interest in profiling to assist with contractual arrangements and with choosing integration partners. We hope this paper will help improve statistical methods for addressing these issues and will encourage the use of interval standards to assess providers.

Much of the profile evaluation literature addresses the difficulties of measuring health care provider quality, including database accuracy, patient and provider confidentiality, and making risk adjustments to handle case-mix differences (Kassirer, 1994). Comparing providers is complicated by differences in hospital volumes and by differences in the risks of the patients they treat. Adjusting for differences should include all variations out of a provider’s control, without adjusting for any variation in its control. Ideal risk adjustment is unrealistic, however, and researchers therefore face a significant challenge (Newhouse, 1996).

For example, profile analyses must account for unequal patient volumes across providers. Garnick et al. (1989) review literature on establishing minimum volume requirements for surgeries and find that the large unexplained variance of provider outcomes makes establishing a minimum volume difficult. Sloan et al. (1986a; b) report an inadequate basis for setting minimum volume standards because the variance of mortality rates was much higher for low volume hospitals than for high volume hospitals. These studies, and others like them, demonstrate the need for better ways to account for unequal sample sizes across providers. Hierarchical models, like the one considered here, can accomplish this goal.

The main example here concerns the US Department of Veterans Affairs (VA), which operates over 170 hospitals as part of one of the largest US health care delivery systems. Intense interest has arisen in performance monitoring and profiling VA facilities because of needs to allocate the VA’s $17 billion budget appropriately (Lehner et al., 1996), to improve accountability of providers insulated from competitive market forces, and to integrate health care service networks within and outside the VA system.

Self-reporting by facilities in the spirit of Continuous Quality Improvement (CQI) and of Total Quality Management (TQM) is desirable, but obtaining consistent reporting patterns from all VA facilities has proved difficult or impossible (Burgess, 1995). Consequently, the VA has developed a set of 15 annual performance monitors that it has used since 1988 to measure the outcomes and processes of its hospitals. Our example focuses on Fiscal Year 1995 for the particular monitor that records the fraction of patients returned to Intensive Care Units (ICUs) at the VA within 3 days of transferring out, if the patient’s ICU return was part of the same admission. This return rate measures the ability of a hospital staff to assess patient needs for continuing ICU care and; ceteris paribus, lower rates indicate better care. A related issue concerns whether reductions in ICU days induced by managed care foster too many readmissions.
Risk adjustments for the return rate monitor are based on average return rates for more than 400 Diagnosis Related Groups (the DRG associated with the original ICU stay) and two age groups (over and under 65), estimated from over 100,000 VA patients nationally. Each ICU’s expected number of returns (denoted $e_i$ for Hospital $i$ in later sections) was computed based on these patient level variables and compared by a statistical test to the ICU’s observed returns. The VA perceived that the result produced an excessive number of outliers, thereby threatening the continued viability of profiling.

The methods offered here can help the VA, and other organizations, to calibrate outliers better and to design better performance standards. The VA might use the results to identify outlying hospitals so as to commend exemplary hospitals and to encourage underperforming hospitals to improve. Whether these measures also should influence budgeting or curtailing programs in hospitals remains an open question, partly because difficulties persist in achieving adequate risk adjustment. Nevertheless, as feedback between information and providers develops, and as profiling methods improve, such uses probably will be considered.

The key contributions here include using interval standards for providers and addressing these standards with probability statements derived from hierarchical models. Section 2 develops the main ideas at a less technical level and, using ICU data, compares them with a commonly used procedure based on $P$-values. Section 3 provides the technical description of the model, while Section 4 explains the statistical method and interprets the numerical results for ICU returns. Section 5 discusses the methods presented in the context of broader health economics research. Section 6 concludes by reviewing the VA’s hospital monitoring system, the advantages of the recommended approach, and some needed extensions.

2. Improving the approach to medical profiling

This section introduces and illustrates two improvements on $P$-value profiles. One improvement employs a hierarchical model to assess inter-hospital information and uses that information to evaluate performances. The other improvement replaces point standards with interval standards.

2.1. A widely-used profiling procedure

Profilers commonly evaluate medical units by testing the hypothesis that a unit’s ‘‘true’’ mean, the hypothetical result of serving infinitely many patients, equals the population average (across all units). A hypothesis often is known to be false a priori because each unit’s true mean will differ, if only slightly, from any pre-specified standard. Further discussion is in the context of profiling ICU returns for the 148 VA hospitals that had ICU units in 1995. The ICU return rate is the fraction of patients that returned within 3 days after release from an ICU.
Interpreting the overall ICU return rate of 2.35% as a precise standard defines hospitals with true return rates of 2.36% as providing “substandard” care (lower rates are better, per Section 1) and those with true return rates of 2.34% as “exemplary”, despite the trivial practical differences. Experienced analysts know that large samples can produce statistical significance without practical significance, but quantifying “practical significance” is challenging (considered further in Section 2.3).

ICU return counts follow Poisson distributions because individual patient return rates are small and outcomes are independent. Medical profilers commonly approximate Poisson distributions with Normal distributions via “Z-scores.” The Z-score for Hospital $i$ is $Z_i = (N_i - e_i)/\sqrt{e_i}$, $N_i$ being the observed ICU return count and $e_i$ the expected number of returns (see Section 1 about computing $e_i$). The central limit theorem states that $Z_\xi \sim \text{Normal}(0,1)$ approximately for large $e_\xi$, so performance on a monitor is “substandard”, “adequate”, or “exemplary” according to $Z_\xi < -1.96$, $-1.96 \leq Z_\xi \leq 1.96$, or $Z_\xi > 1.96$. This amounts to identifying adequate care with a two-sided $P$-value, which exceeds 0.05 because $-1.96$ and $1.96$ are the Normal’s 0.025 and the 0.975 quantiles.

A $P$-value is the frequency in repeated sampling of outcomes no less extreme than the observed outcome, assuming standard performance. Exact $P$-values are computed easily for Poisson data, so the Normal approximation is unnecessary (and it should be avoided). The exact upper tail $P$-value is the probability that a Poisson variable with mean $e_i$ is at least $N_i$, and it also equals the probability that a $\chi^2$ variate with 2 $N_i$ degrees of freedom does not exceed $2e_i$. Thus, tables of the $\chi^2$ distribution can be used for these $P$-value calculations. (In later sections, we refer to the Gamma distribution instead of the $\chi^2$ only because “Gamma” is the more familiar term when assessing prior distributions.) Ulm (1990), Luft and Brown (1993), and Christiansen and Morris (1997a) discuss using Poisson $P$-values in profiling situations.

Table 1 lists ICU return data for four VA hospitals, two $n$s of 459 and 1294 of the 10% more extreme in the substandard direction determined by exact $P$-values with case-mix adjustments, and two $n$s of 418 and 1275 of the 10% more exemplary. The exact Poisson $P$-values ($PV_e$), Column 5, can be calculated from data in Columns 1–3. Thus, the first hospital has $PV_e = 0.138 = P(\text{Poisson}(10) \geq 14) = P(\chi^2_{20} \leq 20)$.

### 2.2. Replacing indirect $P$-values with direct probabilities

Adjustments for hospital level variables are needed if the patient level case-mix is not balanced across all hospitals. The hierarchical model that produced the Column 8 results of Table 1 used covariates that indicate hospital type (Section 4), so to preserve numerical comparability, the $PV_e$ values in Column 5 are adjusted to the corresponding $P$-values $PV_\alpha$ in Column 6. These adjustments otherwise are
Table 1
Comparing profiling methods for four ICU units

<table>
<thead>
<tr>
<th>Raw data</th>
<th>Adjusted for hospital type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>n&lt;sub&gt;i&lt;/sub&gt;</td>
<td>N&lt;sub&gt;i&lt;/sub&gt;</td>
</tr>
<tr>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Possible Substandard Hospitals</td>
<td></td>
</tr>
<tr>
<td>459</td>
<td>14</td>
</tr>
<tr>
<td>1294</td>
<td>42</td>
</tr>
<tr>
<td>Possible Exemplary Hospitals</td>
<td></td>
</tr>
<tr>
<td>418</td>
<td>4</td>
</tr>
<tr>
<td>1275</td>
<td>26</td>
</tr>
</tbody>
</table>

irrelevant to Section 2, and the adjusted P-values PV<sub>A</sub> in Column 6 obey the same logic as the P-values in Column 5.

Define ρ<sub>i</sub>, the “relative” ICU return rate at Hospital<sub>i</sub> (for a standardized mix of patients), as the ratio of that hospital’s true return rate to the mean return rate for all hospitals with the same hospital level covariates. A hospital with ICU returns at the population mean (0.0235) would have ρ = 1.00, assuming average patient case-mix and covariates. The P-values in Column 6 equal (mathematically) the posterior probabilities in Column 7 that ρ<sub>i</sub> for Unit<sub>i</sub> exceeds (first two ICUs of Table 1) or falls below (latter two ICUs) 1.00, where Bayes calculations with an “uninformative” prior distribution for ρ<sub>i</sub> determine the Column 7 entries. (This “uninformative” prior distribution gives ρ<sub>i</sub> a χ<sup>2</sup> distribution with a very large coefficient of variation (CV); equivalently, it puts a uniform distribution on log(ρ<sub>i</sub>).) Column 7 is labeled “CV = ∞” because this uninformative prior distribution is appropriate if ICUs differ sufficiently to justify an infinite coefficient of variation for the random effects ρ<sub>i</sub>.

Columns 6 and 7 agree numerically, but their interpretations differ. P-values, like those of Column 6, are tail probabilities about observed data. Because they are not meant to be probabilities about ICU performances, they can only provide “indirect” information about acceptable ICU performance. Column 7 makes “direct” probability statements about the likelihood of each unit’s actual performance if the CV of the distribution for the ρ<sub>i</sub> values greatly exceeds their estimated standard errors. A large CV for the ICUs is not medically reasonable, and to support this, the 148 observed ICU return rates vary only from 0% to 4.3%, with the CV (weighted for sample size) of the observed return rates being just 0.37. The hierarchical model (Sections 3 and 4) more accurately estimates the CV of the true return rates as 0.205, so the standard deviation of the true ρ<sub>i</sub>s is but 0.5%, or about 20% of their mean (0.0235). The true values ρ<sub>i</sub> lie much closer to the mean of all hospitals than do the observed values. It follows that the probabilities reported in Column 7 are too small, and that the P-values (PV<sub>A</sub>) in
Column 6 similarly exaggerate the evidence that the ICUs did not perform adequately.

Column 8 corrects these biases, “CV estimated” indicating that the CV of the prior $\chi^2$ distribution was set at 0.205. Because 0.205 is the CV of a $\chi^2_{48}$ distribution, the Column 8 values may be viewed as Bayes calculations with a prior $\chi^2$ distribution for $\rho$ (but dividing by 48 to make $E(\rho) = 1.00$ a priori). These prior distributions were estimated by the data, so their reliance on guesswork is minimized, and the Poisson hierarchical model is known to provide excellent repeated sampling (frequency) properties (Christiansen and Morris, 1997a).

The value 0.146 (Column 8) for the small ICU in the third hospital in Table 1 is about five times larger than Column 7, with the precise corresponding odds ratios for Columns 8 and 7 (or 6) shown in Column 9. For example, the odds ratio for this ICU is $5.31 = (0.146)/(1 - 0.146)/(0.031)/(1 - 0.031)$. We interpret this to mean that the third ICU had a good chance of returning patients to ICUs less often than average (85%), but not an overwhelming chance (97%). The other three hospitals in Table 1 also had values (Columns 6 and 7) that overstate the probabilities of being extreme, with greater disparities occurring for small hospitals. This anti-conservative bias in P-values is a general phenomenon in hierarchical model settings (Morris, 1987). The bias may be neglected when samples are large enough to make the within-unit standard errors negligible relative to the between-unit coefficient of variation, but many profiling applications will lack sufficiently large within-unit samples.

2.3. Interval standards

In practice, adequate care corresponds to an interval or range of performance, and we recommend that standards be specified with two values $\rho_L < 1.00 < \rho_U$ (L and U mean lower and upper). Monitors then provide exemplary care if $\rho_I < \rho_L$, and substandard care if $\rho_U < \rho_I$ (assuming that high values indicate poor performance, as for the ICU monitor). Choices of $\rho_L$ and $\rho_U$ reflect many considerations, including medical advances, costs, health consumer needs, and social values. They should be agreed on by a range of knowledgeable and interested parties. The appropriate choices cannot be determined statistically, even though statistical analyses of past performances will provide valuable information. We offer no recommendation for $\rho_L$ and $\rho_U$, and so illustrate the approach for various intervals.

The direct probabilities advocated in Section 2.2 combine easily with interval standards. Two probabilities concerning $\rho_I$ can be calculated for each ICU, after fitting the hierarchical model to all the ICUs. Had the values $\rho_L = 0.8$ and $\rho_U = 1.2$ been chosen, for example, the probabilities would be calculated that each hospital’s true ICU return rate (appropriately adjusted for case-mix and hospital covariates) was less than 80% and was more than 120% of the population average. The horizontal axes in Fig. 1 include ratios $\rho_I$ between 1.00 and 1.60 as possible.
Fig. 1. Illustrating that performance standards affect the probabilities of extreme performance. The hospitals from Table 1 are used here to illustrate how performance standards affect profiling results. The figure on the left (Fig. 1A) shows the probability of substandard performance for two hospitals for criteria values $r$ of 1.0 to 1.6. The probability that $r$ is less than 1.0 for the hospital with 459 patients is $0.74 - 0.26$, which is the probability reported in Table 1, Column 10 from the hierarchical modeling approach. The figure on the right (Fig. 1B) shows the probability of exemplary performance for two other hospitals for a range of $r$ values.

Substandard care standards and ratios $p_L$ between 0.60 and 1.00 as exemplary care standards. The vertical axes give the probabilities of substandard care in Fig. 1A for the two possibly substandard hospitals of Table 1 and of exemplary care in Fig. 1B for the two possibly exemplary hospitals. Any value set for substandard care with $1.0 < p_U < 1.20$ makes the larger hospital ($n = 1294$) in Fig. 1A be a more likely candidate for substandard care than the smaller, with reversal if $p_U > 1.20$. Probabilities for the choices $p_U = 1.0$ and of $p_L = 1.0$ are 1.00 minus the probabilities in Column 8 of Table 1 (see Fig. 1 caption). The substandard care probabilities diminish as $p_U$ increases, and more rapidly for the large $n_l$ hospitals, as Fig. 1A illustrates. Analogous interpretations hold for exemplary care assessments in Fig. 1B.

Because every unit will have some chance of exceeding any chosen standard, threshold probabilities are needed for substandard and for exemplary care probabilities. A natural choice of threshold, although not compelling, is 50%, because that makes units more likely than not to have been substandard (or exemplary) be
designated thusly. The four hospitals in Fig. 1 have probabilities above 50% only if $p_L$ or $p_U$ is close to 1.00. The small hospital in Fig. 1B would be exemplary by this 50% standard only if $p_L$ were between 0.80 and 1.00 (several of the other 144 ICUs were more extreme, however).

In summary, this section states that common profiling methods cause excessive outlier identification because $P$-values generally overstate evidence against null hypotheses, and because interval standards are not used. It recommends using hierarchical models to evaluate directly relevant probabilities and to evaluate the probabilities of adequate care as defined by interval standards.

3. Hierarchical models as an approach

This section further discusses the hierarchical models introduced in Section 2.2. As described in Section 3.1, these models account for regression-to-the-mean, unequal sample sizes, and risk adjustment. Section 3.2 details the specific Poisson hierarchical model needed for the ICU data analyses in Section 4. Hierarchical models with other distributional assumptions (e.g., Normal or Binomial) are available for use, with health care examples in Goldstein and Spiegelhalter (1996), Normand et al. (1997), and Christiansen and Morris (1997b).

3.1. Hierarchical models

Although the ideas in this section are general, for ease of exposition, we discuss them in terms of the ICU returns example. Hierarchical models account for regression-to-the-mean by providing estimates of true hospital ICU standardized return ratios (SRRs) that are appropriately less extreme than observed ICU SRRs. Hospitals with small sample sizes are more likely to have extreme observed ICU SRRs because of chance variation, their true SRRs usually being less extreme than their observed ratios (Morris and Christiansen, 1996). Hierarchical models use data from all the hospitals to make estimates for each hospital by “shrinking” the observed SRR for an individual hospital toward an average ratio. Risk adjustment is incorporated by specifying that this average ratio depend on covariates via a regression structure. “Shrinkage factors” ($B_i$ for hospital $i$) that estimate the amount of regression-to-the-mean are defined in Section 3.2. The use of shrinkage estimates often leads to “crossover” patterns, so that the rankings of the estimated true SRRs from a hierarchical model may differ from those determined by the observed ratios. Estimates from hierarchical models provide more accurate assessments, with the most improvement for smaller hospitals because they experience greater regression-to-the-mean. A benefit of analyses using hierarchical models is that hospitals with small caseloads need not be dropped, whereas other profiling
methods usually omit such units (see, for example, Iezzoni et al., 1994; Welch et al., 1994; and Tucker et al., 1996).

3.2. A Poisson hierarchical model

The Poisson model is appropriate for ICU return counts because the return rates are small, averaging 2.35%, and because the total number of patients returned to the ICU is a sum of independent (possibly non-identically distributed) Bernoulli (zero–one) observations, each with a small probability.

An intermediate goal of the statistical analysis is to obtain distributions for each hospital’s true ICU SRR \( \lambda_i \). The approach here is that of Christiansen and Morris (1997a). For the observed outcome \( N_i \), representing the number of ICU returns within 3 days of discharge at facility \( i \), the model is:

\[
N_i | \lambda_i \sim \text{Pois}(e_i \lambda_i), \quad \text{independently,} \quad i = 1, \ldots, k
\]  

at Level I, where \( k = 148 \), the number of hospitals in the sample. Here, \( e_i \) is the expected number of patients discharged from the ICU at facility \( i \), and \( \lambda_i \) is the expected value of the ratio of the number of patients returned to the ICU to the expected number of returns. The observed ratio, \( y_i = N_i / e_i \), has mean \( \lambda_i \) and variance \( \lambda_i / e_i \).

At Level II, the unknown ICU SRRs \( \{ \lambda_i \} \) vary among hospitals and are assumed to follow independent Gamma distributions with means \( \mu_i = \exp(x_i^T \beta) \) and variances \( \mu_i^2 / \zeta \), for \( \beta \) an \( r \)-dimensional vector. (These Gamma distributions play the same role as the \( \chi^2 \) distributions in Section 2.2.) The mean \( \mu_i \) depends multiplicatively on the covariates \( x_i \). The ratio of the true SRR to its expected value, \( \rho_i \equiv \lambda_i / \mu_i \), then follows a Gamma \( (\zeta, 1) / \zeta \) distribution, having unit mean and squared coefficient of variation \( 1 / \zeta \). In this formulation, the \( k \) facilities are “exchangeable” in the sense that all of the \( \rho_i \) follow the same distribution (more generally, exchangeability requires that their joint distribution be invariant under all permutations). In summary, the Level II distributions are:

\[
\rho_i | \zeta, \beta \sim \text{Gamma}(\zeta, 1) / \zeta, \quad \text{independently,} \quad i = 1, \ldots, k.
\]

We follow Christiansen and Morris (1997a) and carry out a Bayesian analysis by assuming little prior information about the unknown \( (\zeta, \beta) \). Specifically, \( 1 / \zeta \) is uniformly distributed on the positive real line, and \( \beta \) is uniformly distributed in \( r \)-space. In simulation studies, this specification for \( \zeta \) and \( \beta \) has been shown to yield posterior intervals that meet their nominal coverages and with lower squared error risk than other methods.

Given \( N_i, x_i \), the hyperparameters \( \zeta \) and \( \beta \), and the models specified in Eqs. (1) and (2), the true ICU SRRs \( \lambda_i, \quad i = 1, \ldots, k \), follow independent Gamma posterior distributions with means:

\[
E(\lambda_i | \text{data, } \zeta, \beta) = (1 - B_i) y_i + B_i \mu_i
\]  

(3)
and with variances

\[ \text{Var}(\lambda_i|\text{data}, \xi, \beta) = \frac{E(\lambda_i|\text{data}, \xi, \beta)}{e_i} (1 - B_i). \]

(4)

In Eqs. (3) and (4),

\[ B_i \equiv \frac{\xi}{\xi + e_i \mu_i}; \quad 0 < B_i < 1. \]

(5)

When \( B_i \), the “shrinkage factor” from Section 3.1, is large (i.e., \( e_i \mu_i \) is small so there is little information about individual facility \( i \) or \( \xi \) is large because the true ratios are very similar), the expected true ratio in Eq. (3) is weighted heavily toward \( \mu_i \) (an average ratio, which in our example is a function of the type of hospital). If \( B_i \) is small (i.e., either \( e_i \mu_i \) is large so there is much information about individual facility \( i \) or \( \xi \) is small because there is much heterogeneity in the data), then the expected true ratio is weighted heavily toward the individual estimate \( y_i = N_i/e_i \). From Eq. (4), a facility with a smaller expected number of returns generally will have a larger proportionate reduction in variance (Morris, 1983a).

To obtain the moments (3) and (4), estimates of \( \xi \) and \( \beta \) are needed. The marginal distribution of the \( N_i \) is used for this purpose, being determined from Eqs. (1) and (2) as:

\[ N_i|\xi, \beta \sim \text{Negative Binomial}(\xi, 1 - B_i). \]

(6)

Technically, this means that \( N_i \) has the same distribution as the number of successes that occur before \( \xi \) failures in a sequence of Bernoulli trials with success probability \( 1 - B_i \). The density in Eq. (6) for \( i = 1, \ldots, k \) and the prior density for \( \xi \) and \( \beta \) allows inferences about \( (\xi, \beta) \), computed by the program PRIMM (Poisson Regression Interactive Multilevel Model) (Christiansen and Morris, 1996b). PRIMM calculates approximate posterior distributions for the true ICU SRRs \( \{\lambda_i\} \) by accounting for the uncertainty in the estimates of \( \xi \) and \( \beta \). If estimates of \( \beta \) and \( \beta \) are substituted into Eqs. (3) and (4) without accounting for the uncertainty in their estimation, interval widths are underestimated, resulting in the illusion of more certainty than is inherent in the problem.

4. Analysis, profile results and model checking

The analyses that follow combine data from all 148 VA hospitals to fit the hierarchical Poisson model presented in Section 3.2. In Section 4.1, we discuss estimating the unknown Gamma distribution parameters in Eq. (2) and the
inclusion of covariates. In Fig. 1, in Section 2, we already have demonstrated how the choice of exemplary and substandard criteria affects conclusions about hospital performance. In this section, we propose and illustrate how all of the information needed by hospital administrators to assess and compare hospitals might be contained in one chart. In Section 4.2, we discuss how we addressed the crucial issue of checking the hierarchical model fit.

4.1. Fitting the hierarchical model

An approximate Gamma distribution is determined for each hospital to describe a parameter \( p_i = \frac{E(y_i|\lambda_i)}{E(y_i|\xi, \beta)} \), that represents the ratio of the expected case-mix standardized ratio to the population ratio. These probability distributions are used with the standards introduced in Section 2.3 to calculate profiles.

Overall, 2.35% of the VA hospital patients returned to ICUs within 3 days. The expected number of ICU returns at Hospital \( i \), \( e_i \), is the sum of the individual return probabilities for the \( n_i \) patients. These \( e_i \) values, provided by VA to make adjustments so hospitals that treated high-risk patients would not be penalized unfairly, are based only on patient-level variables: their ages and DRG categories. The model and PRIMM software expects the counts \( e_i \) to be non-random. In fact, these values were estimated by VA using the full database of 104,000 patients. While this is random, the size of this dataset provides estimates of the \( e_i \)'s that are sufficiently accurate to allay concern.

Additional adjustments are based on hospital-level characteristics. The 148 hospitals are classified into six types (Stefos et al., 1992) using data not presented here. The descriptive names attached to the groups by the Stefos et al. study are listed in Column 1 of Table 2; the number of hospitals by type and each type’s

<table>
<thead>
<tr>
<th>Hospital type (j)</th>
<th>No. of hospitals ( k_j )</th>
<th>Average No. of patients ( n_j )</th>
<th>Average observed rate (%)</th>
<th>Average expected rate based on DRG and age (%)</th>
<th>( \frac{\sum N_i}{\sum e_i} )</th>
<th>( \exp(\beta) ), estimated adjustment to ( e_i ) by type (9% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small teaching</td>
<td>15</td>
<td>442.3</td>
<td>2.00</td>
<td>2.17</td>
<td>0.92</td>
<td>0.93 (0.76, 1.13)</td>
</tr>
<tr>
<td>Midsize teaching</td>
<td>47</td>
<td>952.9</td>
<td>2.53</td>
<td>2.40</td>
<td>1.05</td>
<td>1.04 (0.95, 1.13)</td>
</tr>
<tr>
<td>Metro teaching</td>
<td>26</td>
<td>1321.2</td>
<td>2.67</td>
<td>2.45</td>
<td>1.09</td>
<td>1.08 (0.98, 1.20)</td>
</tr>
<tr>
<td>Small general</td>
<td>32</td>
<td>274.3</td>
<td>1.25</td>
<td>2.12</td>
<td>0.59</td>
<td>0.59 (0.48, 0.72)</td>
</tr>
<tr>
<td>Midsize general</td>
<td>14</td>
<td>551.0</td>
<td>1.81</td>
<td>2.16</td>
<td>0.84</td>
<td>0.80 (0.65, 0.98)</td>
</tr>
<tr>
<td>Psychiatric</td>
<td>14</td>
<td>238.6</td>
<td>1.59</td>
<td>2.08</td>
<td>0.76</td>
<td>0.78 (0.57, 1.03)</td>
</tr>
</tbody>
</table>

Table 2: Hospital types and fitted population parameters. Column 3 is the sum of \( n_i \) within hospital type \( j \) divided by \( k_j \), \( \sum n_i / k_j \); Column 4 is \( \sum \exp(\beta) \); Column 5 is \( \sum \exp(\beta) e_i / k_j \); Column 6 is \( \sum N_i / \sum e_i \); and so is the ratio of Columns 4 to 5. Column 6 shows estimates from PRIMM of adjustments for hospital type.
average caseload are in Columns 2 and 3. The averages (weighted by caseload) within hospital type of observed return rates \( N_i / n_i \) and of expected rates \( \bar{e}_i = e_i / n_i \) are in Columns 4 and 5 of Table 2. We see that the combined 15 small teaching hospitals had 2.00 per 100 of their patients return to the ICU (Column 4), less than their combined DRG and age adjusted expected rates of 2.17% (Column 5) suggest.

VA wishes to compare hospitals within their type because this classification is thought to serve as a proxy for within DRG case-mix severity. For example, teaching hospitals usually treat more severely ill patients within a particular diagnosis than the non-teaching hospitals do. Therefore, adjustments to the expected values are needed based on hospital classifications. Columns 6 and 7 provide two possible adjustments to \( e_i \) due to case-mix differences by hospital type. Column 6 is the standardized ratio of the observed counts to the expected number of returns within a hospital group, i.e., it is the average ratio (weighted by \( e_i \)) for each hospital type. These values were used to standardize ratios by hospital group in the calculations for Table 1, Column 10. Column 7 presents estimates from the hierarchical model as discussed next. Adjusting the number of patients who returned to the ICU based on the expected number of returns, \( e_i \), and then also by hospital type, one can compare all hospitals on the same scale. This amounts to assessing the \( \rho \).

In this example, the regression (2) is specified with a 148 by six dimensional matrix \( X \); the six columns indicate the six hospital types using 0s and 1s. The observed ratio \( y_i = N_i / e_i \) has marginal expectation:

\[
E(y_i | \xi, \beta) = \mu_i = \exp(x_i^T \beta).
\]  

Because \( x_i \) indicates hospital type, all \( k_j \) hospitals of type \( j \) have the same expectation. The hierarchical model estimates of these adjustments to \( e_i \) (Column 7), of course, are close to the ratios in Column 6 because they estimate the same quantities. The values of \( \exp(\hat{\beta}_2) \) and \( \exp(\hat{\beta}_3) \) of 1.04 and 1.08 in Column 7 for midsize and metro teaching hospitals mean that these types of hospitals had 4% and 8% more ICU returns than their patient characteristics suggest if only \( e_i \) is used. Accordingly, the final expected ICU return for the \( i \)th hospital, assuming average care for its patient mix and hospital type, is the product of \( e_i \) and the adjustment from Column 7.

VA also considers adjustments for other hospital characteristics, such as the number of beds and the occupancy rate, because these variables could attract different patient-mixes. For this monitor, these variables were analyzed and found to be statistically insignificant. A warning about adjustment variables: Hospital-level adjustments are justified if they help identify unmeasured patient characteristics that are associated with the outcome. They are not justified for hospital-level variables associated with the quality of care if the measured aspect of quality of care is under the hospital’s control. For example, if sub-optimal staffing adversely affects quality of care, a profile analysis adjusting for a staffing variable would
forgive hospitals with poor staffing practices. However, it still could be important to estimate this source of performance variation.

PRIMM provides the estimate $\xi = 23.9$ with a 95% interval (13.0, 43.8). Thus, the estimated coefficient of variation is $1/\sqrt{\xi} = 1/\sqrt{23.9} = 0.205$, the value used in Section 2.3. The (population) standard deviation of each $\lambda_i$ is about 20.5% of $\lambda_i$; similarly for the $\rho_i$ values. The 95% interval for this coefficient of variation is (0.151, 0.277), so 27.7% is an upper bound for how much the true ICU return rates vary at the hospitals. In other words, hospitals are fairly homogeneous. In contrast, the model in Column 9 of the table in Section 2 assumes $\xi$ is infinite and the coefficient of variation is 0.

Each hospital’s estimated expected ratio, $\hat{\lambda}_i$, is a weighted average of its $\hat{\mu}_i$ and its observed ratio $y_i = N_i/e_i$. Our inference is then based on the distribution of $\rho_i = \lambda_i/\mu_i$. The estimated shrinkage factors $\hat{B}_i = \xi/(\xi + e_i \hat{\mu}_i)$ average 0.65 and range from 0.31 (a hospital with a large adjusted expected count) to 0.97 (a hospital with a small adjusted expected count).

Fig. 2 displays the profiling results for some of the most exemplary and some of the most underperforming hospitals among the 42 in the Eastern region. This chart could be used to show individual hospital administrators how their hospital performed on this monitor compared to others in the region. The rows are ordered from top to bottom by increasing values of $\hat{\lambda}_i$. Each row represents one hospital and provides the number of patients released from the ICU during fiscal year 1995, $n_i$, and its estimated shrinkage factor, $\hat{B}_i$, as fitted by the hierarchical model.

Each horizontal line in Fig. 2 is the 95% probability interval for the standardized true return ratio $\rho_i$ for that hospital. That is, each interval contains the true $\rho_i$ with probability 0.95, adjusted for everything known. Arbitrary standards for exemplary performance of (0.80) and for poor performance of (1.20) are illustrated by the two corresponding vertical lines. The values listed next to these vertical lines are the probabilities that the true $\lambda_i$ values were in the exemplary range (number near left end) and in the underperforming range (number near right end). Note that the probabilities for the illustrated performance criteria are not monotone. The variances depend on shrinkages and caseload which differ among the hospitals. Three of the four hospitals in Fig. 1 are identified by an asterisk on the left side of the figure. A crude estimate of $\rho_i$ is the ratio of the observed $y_i$ to the regression estimate $\hat{\mu}_i$, and is marked with ($\triangle$). The hierarchical model estimate of the true ratio, $\hat{\rho}_i$, is marked with ($\cdot$).

Rankings of hospitals by their $P$-values and by hierarchical models often differ. To illustrate one rank comparison for ICU data, we use ranks of the $P$-values defined by the probability that $\rho_i > 1$ using the model with an infinite CV (as discussed for Column 9 of Table 1) and ranks using the hierarchical model estimates of the probability that $\rho_i > 1$ (as discussed for Column 10 of Table 1). Thirty percent of the 148 hospitals changed by at least 10 ranks, the largest change
Fig. 2. Charting hospital performance. This chart shows nine of the most exemplary and nine of the most underperforming hospitals out of the 42 VA hospitals in the East. They are ordered by $\tilde{\rho}$, which is represented by $\Delta$. The $\Delta$ identifies $y_j/\mu_j$. Each horizontal line is the 95% probability interval for the true performance ratio $\rho_i$ in the corresponding hospital. The dashed vertical lines indicate two possible performance criteria for exemplary and substandard care. The numbers next to these lines are the probability that $\rho_i$ is less than (on the left) and greater than (on the right) the chosen criteria. For example, for the hospital third from the bottom, the probability that its true ratio $\rho_i$ is less than 0.8 is 0.04. The probability that $\rho_i$ is greater than 1.2 is 0.36. The number of patients $n_i$ and the estimated shrinkage factor $\tilde{B}_i$ are given on the right. Hospitals also used in Table 1 and Fig. 1 are marked with an asterisk.

being 32 ranks. For theoretical reasons, the differences at the most extreme ranks are relatively limited, but in one case the ICU with the 13th highest $P$-value (indicating substandard performance) had only the 37th highest hierarchical model rank. Of course, these comparisons are for the disfavored standard ($\rho_i > 1.00$), see Section 2. With the preferred interval performance criteria, the hierarchical model ranks usually would differ even more from $P$-values. But whether they alter the relative rankings of hospitals or not, hierarchical models are needed mainly to provide directly interpretable probabilities and to make use of appropriate perfor-
mance standards. Together these enable administrators to make better decisions about hospital performance.

4.2. Checking the hierarchical model

Both levels of the hierarchical model assumptions in Section 3.2 need to be considered carefully based on the data at hand, experience with related data sets, and theory. Techniques for the Level I model checks now are widely understood, but Level II checks, which are for hierarchical models, are relatively new.

For our data, the Level I Poisson distribution (1) holds for theoretical reasons, as stated in Section 3.2. The Gamma distribution at Level II is selected partly because it produces robust inferences. That is, the Gamma distribution minimizes the maximum mean-squared-error for estimating the true hospital return ratios $\lambda_i$ (Morris, 1983b). Even if the Gamma distribution is the wrong choice, the estimates of the means and variances still remain valid. No other Level II distribution for Poisson data provides this robustness.

Our approach to checking the Level II assumptions (Eq. (2)) in a hierarchical model starts with the assumption that the Level I distribution is valid. Then, any departure of the observed data $N_i$ from their marginal Negative Binomial distribution (Eq. (6)) could only be due to an incorrect specification at Level II. We made significance tests and used graphical methods to compare the distribution of the observations $N_i$ with the Negative Binomial distribution (Eq. (6)). When additional terms are entered into the regression model (7) (e.g., transformations of the $n_j$ values and of the other hospital variables made available to us by VA), the significance of these new regression coefficients was marginal.

We plotted the standardized residuals, $r_i = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\sigma}^2 + \hat{\mu}_i}}$, against various hospital-level predictors. Standardized residuals are approximately independent deviations of the $N$ vector from its expectation, measured in standard deviation units, i.e., they have means of zero and unit variances. The residuals showed patterns similar to those expected when sampling independently from the appropriate Negative Binomial distributions.

A key Level II assumption, usually the hardest to test, is that the $\rho_i = \lambda_i / \mu_i$ values are identically distributed (have exchangeable distributions), as in Eq. (2). The residual plots against the hospital-level variables support this if the residuals have the same distribution in all vertical sections of the plot. This also was tested formally with a $\chi^2$ test, which revealed no significant differences among estimates of $\zeta$ made for the six separate hospital types. These Level II checks for the Poisson setting are described more fully in Christiansen and Morris (1996a).

5. Discussion

We have chosen to provide a finely honed example of recommended profiling methods in this paper that illustrates the gains obtained by deriving probability
statements from hierarchical models using interval standards. Nevertheless, the methods discussed here have much wider implications and possibilities for health economists facing a variety of other healthcare market issues. We briefly present two of those application issues here for discussion. The first issue highlights a fundamental disconnect between theoretical efforts to identify the market effects of different product qualities with empirical efforts by health economists that assume all providers produce the same quality of healthcare services. The second issue highlights further work to be done modeling individual level differences in the context of hierarchical models.

Theoretical industrial organization models on quality differences leave a wide opening for new approaches to empirical work on provider quality differences. Recent models show that dissimilar product qualities can be associated with more intense price competition (Ma and Burgess, 1993) and with less intense price competition (Shaked and Sutton, 1982, 1983) depending on whether exogenous horizontal product differentiation exists. In the presence of horizontal factors such as spatial location, vertical product differentiation (product quality differences) is not required to relax Bertrand competition. A first step for health economists in establishing empirical tests of competing strategic hypotheses is to develop ways of identifying the presence or absence of provider performance variation statistically. Unfortunately, no simple metric exists for measuring performance differences, and identification only can be made probabilistically. This measurement problem is under-recognized in building health care market models that assume all providers produce the same health service quality.

The particular example developed here does not break new ground in using risk adjustment measures in profiling at the individual level. Nevertheless, building hierarchical models that address regression-to-the-mean is extremely important when there is a large amount of variation not explained by regression models at the provider and individual levels. Such incomplete risk adjustment does not appear to be easily solved in practice. Though this work does not incorporate individual level observations directly in the model, further research that directly accounts for individual level variation in this context is ongoing by members of this research team and has been explored to some extent by others (Thomas et al., 1994; Normand et al., 1997).

6. Conclusions

Since 1988 VA has been building a national information system with process and outcome quality monitors, like this ICU monitor, accessible to managers and researchers throughout the VA system. This system includes the observed and expected frequency counts, at-risk sample sizes, and probabilities of high and low extremes for each monitor (as determined by hierarchical models similar to the one described here). In the case of high extremes, this information system and its
abstracts are used by individual medical centers to identify areas that need Continuous Quality Improvement resources most, while the low extremes are used to identify exemplary providers from which to learn.

Besides expanding the scope of problems to which hierarchical models apply, incremental research is needed in several areas, partly to integrate annual data and multivariate data. Work on time series models is needed to account for significant performance changes over time, e.g., akin to that of West and Aguilar (1997) done for this VA database. When many monitors are collected simultaneously for each hospital in a given year, accounting for their correlations also should improve decisions, especially for smaller providers.

This paper recommends basing profiles on hierarchical models and on interval criteria for several reasons. First, the impetus from VA for the hierarchical modeling approach was the excessive outlier identification, most alarmingly of negative outliers, that were produced by using $P$-value methods and sharp null hypotheses. Differences in simple observed/expected rates varied widely from year to year, and undermined the trust of VA clinicians and managers in the measurement process. Second, managers, medical practitioners, and policymakers that we have interacted with have felt comfortable interpreting the probabilities of performances derived from hierarchical models. Conversely, they have expressed discomfort with the interpretations of $P$-values for hospital performances from this data, especially given the large numbers of observed high outliers identified by that method. Third, although profiling with medically meaningful interval standards is handled easily with hierarchical models, the context is unfavorable for using $P$-values and seems not to have been attempted in print. Finally, replacing current implicit standards with requirements for interval standards also could help foster an overdue national debate on standards for medical profiling. In this way, all participants in the healthcare system can meaningfully compare providers.

Acknowledgements

The especially detailed comments of the editors as well as comments from two anonymous referees are gratefully acknowledged. Dr. Morris gratefully acknowledges the support of NSF grant DMS-9705156. The conclusions and opinions in this paper are those of the authors and do not necessarily reflect official positions of the US Department of Veterans Affairs.

References


