CHAPTER

12

COMPARING OUTCOMES ACROSS PROVIDERS

Arlene S. Ash, Michael Shwartz, and Erol A. Peköz

Risk adjustment facilitates meaningful comparisons of outcomes across groups of patients by accounting for those differences in intrinsic patient characteristics that are related to outcomes. However, such comparisons are just a means to a larger end. Nightingale and Codman viewed comparing outcomes as a powerful way to motivate improvement of quality of care (see Chapter 1). As Nightingale wrote in 1863 (175–76):

> In attempting to arrive at the truth, I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any purposes of comparison. . . . I am fain to sum up with an urgent appeal for adopting . . . some *uniform* system of publishing the statistical records of hospitals. There is a growing conviction that in all hospitals, even in those which are best conducted, there is a great and unnecessary waste of life . . .

Nightingale and Codman argued that simply comparing rates of events was insufficient. One must discover why differences in patient outcomes occurred and correct identified problems.

Jumping to the start of the twenty-first century, comparisons of outcomes are now central to scrutiny of the American health care delivery system and an important component of responses to competitive market forces. Patient outcomes are compared across individual physicians, group practices, clinics, hospitals and other institutional settings (e.g., nursing homes), and private and public health insurers. These comparisons are variously called performance or practice profiles, report cards, scorecards, and outcomes reports. As noted in Chapter 1, the growing interest in pay-for-performance schemes will likely put such profiles center stage, heightening the financial stakes of risk-adjusted outcome measures (Institute of Medicine 2002c).

Several types of questions motivate report card and profiling initiatives, such as:

- Do any particular providers stand out as either much better or worse than average?
- How strong is the evidence that provider A's performance has been (or, perhaps more importantly, will be) substandard?

Numerous decisions are required when designing a profiling approach and assembling the data to compare patient outcomes across providers and answer such questions (Table 12.1). In addition, interpreting results requires both good methodology and a thoughtful conceptual framework. This chapter discusses "principles of good design" as well as important practical considerations in performance profiling. We emphasize, however, that no all-purpose best way exists to compare patient outcomes across providers. Especially when using profiles to support decisions with serious patient care or financial implications, analysts must remain aware of how various methodological choices can shape their findings.

The Effect of Randomness on Comparing Patient Outcomes

Random fluctuations affect estimates of provider performance and thus limit the conclusions that can be drawn safely from performance profiles. To elucidate the role of randomness, we consider a contrived example targeting hospital costs as our outcome. We assume that the patients have identical clinical conditions and receive the same treatment across hospitals.

The simplest model views the costs of the n_A cases admitted to hospital A this year as a sample from a theoretically infinite population of cases that might be treated at hospital A. This year's observed average cost at hospital A, $\overline{Y}_A$, estimates the underlying average cost, μ_A, for all cases that might be treated there.

The distribution of observed costs for this year's patients provides information about the variability of costs among potential groups of patients at various hospitals. Examining this distribution is always advisable. For example, analysts should look at standard summary statistics: the mean, median, and standard deviation; minimum and maximum values; and values associated with different percentage points of the distribution (e.g., the value demarcating the upper 1 percent of cases). For facilities in hospital A's comparison group, side-by-side boxplots (sometimes called box-and-whiskers plots, discussed later; Figure 12.1) help to identify likely errors (e.g., hospital stays with negative costs) or values that are correct but extreme (Tukey 1977). For instance, a hospital with high average costs because all its cases were expensive differs from an institution where one very expensive case raised the average by nearly $10,000 (e.g., one "million-dollar baby" among 100 average-cost newborns).

The standard deviation (SD) is the most common summary measure of variation for a variable Y. It is estimated for a population from a sample Y_1, $Y_2, \ldots, Y_n$, by:

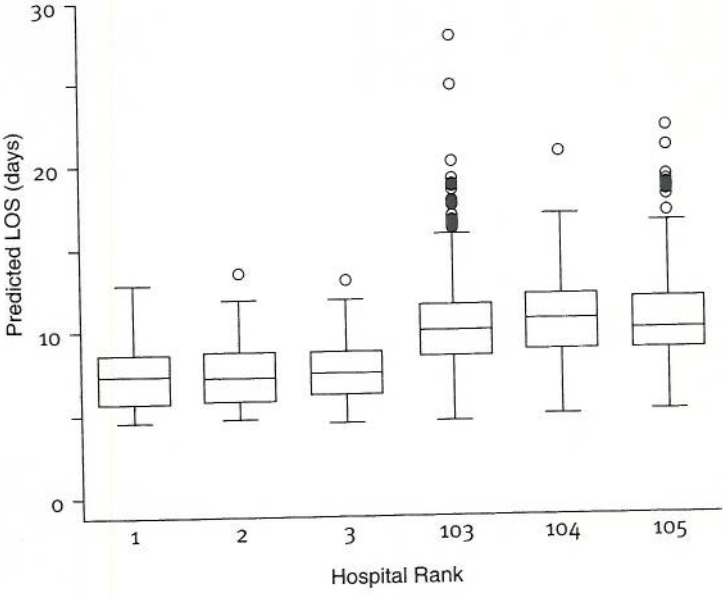
$$s = \sqrt{\Sigma_i (Y_i - \overline{Y})^2/(n-1)}$$

TABLE 12.1
Design Considerations for Provider Profiling

What data will be used?
  Can information be linked at the person level?
  Can numerators and denominators be determined?
  What are the accuracy and reliability of the data?
  Which patient risk factors are captured in the data?
  What is the time frame encompassed by the data?
What outcomes can be measured from the data?
Which providers will be included?
  Are there reasons to exclude any providers?
    - Small sample sizes
    - Incomplete data
    - Known patient risks unable to measure with the data (e.g., public hospitals)
    - Policy considerations (e.g., small hospitals, rural hospitals)

Which patients will be included?
  What are the specific inclusion criteria (e.g., disease, surgery)?
  Are there reasons to exclude any patients?

FIGURE 12.1
Box Plots of Expected LOS at Six Hospitals



NOTE: Box plots from three hospitals with the lowest and three with the highest expected LOS using Disease Staging's Relative Resource Scale to determine expected values.

Just as $\overline{Y}_A$ is an estimate of an unobserved mean, μ_A, of a larger theoretical population, the SD in hospital A, s_A, is an estimate of σ_A, the SD of that larger population. SDs have the same units as Y (here, dollars). Regardless of the shape of the distribution of Y, most costs are likely to lie within two or three SDs of the average.

The usual model hypothesized when comparing hospitals (or any other provider unit) assumes that

- each case at hospital A has expected value μ_A, which may differ by facility; but
- the Y values at each facility are equally variable (i.e., there is a single, common value σ of the σ_As).

If all the SD_As have the same value, σ, each s_A (computed by applying the above formula to the cases at hospital A) is an estimate of it. The "pooled" estimate of s is calculated from the s_As using weights that reflect differences in sample sizes, as:

$$s = \sqrt{\Sigma_A (w_A * s_A^2)}$$

where A indexes facilities and $w_A = (n_A - 1)/(N - K)$, with N being the total number of patients and K the number of facilities.

A common goal is both to estimate μ_A and measure the accuracy of these estimates based on what we observe, namely $\overline{Y}_A$, n_A, and s_A. For almost any distribution of Y, the observed average cost at hospital A, $\overline{Y}_A$, is very likely to fall within two standard errors (SEs) of the true average μ_A—that is, within the range $\mu_A \pm 2 * SE_A$, where SE_A, the (observed) SE of the mean at hospital A, is calculated as:

$$SE_A = s/\sqrt{n_A}$$

Thus, each interval $\overline{Y}_A \pm 2 * SE_A$ is likely to contain its true hospital mean value, μ_A.

For a given hospital A, μ_A (the value around which individual Y values are centered) and σ_A (a measure of how much the individual values are dispersed around μ_A) are properties of the population and do not depend on n_A, the number of cases at hospital A. In contrast, SE_A relates to the variability of an "estimator"; specifically, it measures how accurately $\overline{Y}_A$ estimates μ_A. As the number of observations at hospital A increases, SE_A decreases, reflecting the increased accuracy of $\overline{Y}_A$ as an estimate of μ_A. For example, suppose that we observe mean costs of \$5,000 with an SD of \$5,000.[1] With 100 observations, we are reasonably sure that μ_A will be in the interval $\$5{,}000 \pm 2 * 500$, that is, from \$4,000 to \$6,000. With $n = 400$, we are reasonably confident that μ_A is between \$4,500 and \$5,500.

Even with the highly skewed distributions typical of health care cost data (see Chapter 10), unless n_A is small, the observed mean will be approximately normally distributed. This implies that an interval centered at $\overline{Y}_A$ and extending for two or three SE-sized units above and below will likely include μ_A. For many purposes, 30 cases is an adequate n_A. However, especially when the outcome variable is extremely skewed, producing a nearly normal distribution of $\overline{Y}_A$ may require hundreds of cases, as extreme values significantly affect the mean. When the underlying variable in the population from which the n_A cases were sampled is distributed normally, $\overline{Y}_A$ has about a 95 percent chance of falling in the so-called "prediction interval," $\mu_A \pm 2 * SE_A$. This is

equivalent to saying that the interval $\overline{Y}_A \pm 2 * \mathrm{SE}_A$ has a 95 percent chance of containing the true value, μ_A. This latter interval is called the 95 percent confidence interval (CI). A CI specifies plausible boundaries for a parameter estimate (here, μ_A), whereas a prediction interval establishes boundaries within which an observed random variable (here, $\overline{Y}_A$) should lie. Although we know that cost distributions are not normal, cost averages, such as $\overline{Y}_A$, are more nearly so.[2]

Profiling is not a theoretical exercise in statistical science. Rather, these intervals provide a convenient "cutoff" for highlighting situations requiring further examination. A two-SE cutoff casts a broad net for identifying possible problems, many of which will be spurious. Three SE's might prove more appropriate for settings, such as public reporting, in which false flags have severe consequences.

The observed coefficient of variation (CV) is the ratio of the SD to the mean: $s/\overline{Y}_A$. Table 12.2 shows the half-widths of (approximate) 95 percent CIs:

$$2 * \mathrm{SE} = 2 * \overline{Y}_A * \mathrm{CV}/\sqrt{n_A}$$

for different values of CV and sample size. The figures in Table 12.2 demonstrate the combined effect of sample size and data variability on the range containing the average of a sample of size n. For example, with CV = 1 and a sample size of 100, the approximate 95 percent CI for μ_A is:

$$(0.8 * \overline{Y}_A, 1.2 * \overline{Y}_A)$$

That is, we can estimate mean cost with about 20 percent error. Because accuracy is proportional to 1 over the $\sqrt{n}$, achieving estimates with 10 percent error requires 400 observations.

From a different perspective, assume that average costs are the same at each hospital. If μ_A equals μ, $\overline{Y}_A$ is likely to be within the interval $\mu \pm 2 * s/\sqrt{n_A}$.[3] Assuming a large number of patients across all hospitals, $\overline{Y}$ calculated from all patients is a good estimate of μ, the true mean cost for all patients. Hence, $\overline{Y}_A$ is likely to be within the interval $\overline{Y} \pm 2 * s/\sqrt{n_A}$. Suppose $\overline{Y}_A$ falls outside this interval. In the traditional hypothesis-testing framework, we conclude that hospital A's costs differ from average. Furthermore, suppose that we are judging 100 facilities and we flag any hospital as an outlier when $\overline{Y}_A$ is outside the interval $\overline{Y} \pm 2 * s/\sqrt{n_A}$. In this situation, hospitals designated as outliers have costs that are statistically significantly different than average at $p < 0.05$. Among 100 identical facilities (i.e., all $\mu_A = \mu$), this approach incorrectly flags about five outliers. This is an example of the "multiple comparisons" problem (Snedecor and Cochran 1980). Incorrectly flagging a typical hospital is called a type I error.

On the other hand, trying to avoid type I errors by being conservative about flagging outliers increases the chance that true outlier hospitals are missed—a type II error. To illustrate type II errors, assume that an inefficient

TABLE 12.2 Effect of Sample Size and Coefficient of Variation (CV) on the Half-Width of Approximate 95 Percent Confidence Intervals*

Sample Size (n)	CV (σ/μ)					
	0.5	1	1.5	2	2.5	3
10	0.32	0.63	0.95	1.26	1.58	1.90
25	0.20	0.40	0.60	0.80	1.00	1.20
50	0.14	0.28	0.42	0.57	0.71	0.85
100	0.10	0.20	0.30	0.40	0.50	0.60
150	0.08	0.16	0.24	0.33	0.41	0.49
200	0.07	0.14	0.21	0.28	0.35	0.42
300	0.06	0.12	0.17	0.23	0.29	0.35
400	0.05	0.10	0.15	0.20	0.25	0.30
500	0.04	0.09	0.13	0.18	0.22	0.27
1000	0.03	0.06	0.09	0.13	0.16	0.19

* Cells of the table are $2 * \text{CV}/\sqrt{n}$. Half-width = $\overline{Y}_A *$ table cell.

hospital (hospital I) has costs 20 percent above an average of $1,000. This difference seems sufficiently large to be important. Suppose that we flag a hospital as an outlier only when its observed mean differs from $1,000 by at least 40 percent (roughly the cutoff for identifying a statistically significant difference at $p < 0.05$ when $n = 100$ and the CV = 2). Under this rule, hospital I will be flagged if its average cost exceeds $1,400 (i.e., the 95 percent CI for μ_I lies entirely above $1,000). However, with only 100 observations and a CV of 2, the chance that hospital I's observed mean will exceed this threshold is only 20 percent. Thus, hospital I has an 80 percent chance of avoiding an outlier flag.

The same considerations apply when examining a dichotomous outcome, such as death. Assume that P is the death rate in a large population of patients and $\hat{P}$ is the observed rate in a smaller population of size n (e.g., patients at hospital A). The estimated SE is then $\sqrt{\hat{P}(1-\hat{P})/n}$. For different values of n and death rate P, Table 12.3 shows the half-width of an interval that is about 95 percent likely to contain P.[4] For example, if $\hat{P}$ were 10 percent in a sample of 100 patients, the interval from 4 percent to 16 percent is likely to contain P.[5] This interval is wide compared to reasonable differences between poor- and high-quality providers (Hofer and Hayward 1996; Ash 1996).

With any rule for flagging outliers, as the difference increases between a given hospital's underlying performance and typical performance, the likelihood of being flagged rises. Thus, depending on the (unknown) mix of normal and variously aberrant providers in a study population, roughly 5 percent of nonproblematic providers will erroneously receive outlier flags, whereas some (unknown fraction of) problematic providers will escape flags. Which flags are incorrect is generally not obvious. Using data on cardiac catheterization, Luft and Hunt (1986, 2780) illustrated that small numbers of patients and relatively low rates of poor outcomes make it difficult to "be confident in the

TABLE 12.3 Effect of Sample Size and Probability of Death on the Half-Width of Approximate 95 Percent Confidence Intervals*

Sample Size (n)	Probability of Death (P)**							
	0.01	0.02	0.05	0.10	0.15	0.20	0.25	0.50
25	—	—	—	0.12	0.14	0.16	0.17	0.20
50	—	—	0.06	0.08	0.10	0.11	0.12	0.14
100	—	—	0.04	0.06	0.07	0.08	0.09	0.10
150	—	—	0.04	0.05	0.06	0.07	0.07	0.08
200	—	—	0.03	0.04	0.05	0.06	0.06	0.07
300	—	0.02	0.03	0.03	0.04	0.05	0.05	0.06
400	—	0.01	0.02	0.03	0.04	0.04	0.04	0.05
500	0.01	0.01	0.02	0.03	0.03	0.04	0.04	0.04
1,000	0.01	0.01	0.01	0.02	0.02	0.03	0.03	0.03

* Cells of the table (= half-width) are $2 * \sqrt{P * (1 - P)/n}$.
** When $n * P$ (the expected number of deaths) < 5, the normal approximation (the basis for calculations in this table) is unreliable (—).
NOTE: More precise CIs for proportions are described in Agresti and Coull (1998) and implemented in http://www.graphpad.com/quickcalcs/ConfInterval2.cfm.

identification of individual performers." For example, suppose the death rate is 1 percent, but a hospital treating 200 patients experiences no deaths. Even using a lenient 0.10 significance level, determining whether that hospital had statistically significantly better outcomes is impossible. If the expected death rate is 15 percent and five deaths occurred out of 20 patients (an observed rate of 25 percent), the difference is insufficient to label the hospital as performing poorly. Thus, random chance plays a prominent role in determining outlier status when sample sizes are relatively small. In this situation, comparisons across providers must be interpreted cautiously.

Comparing Observed and Expected Outcomes

Calculating expected rates of outcomes is usually the first step in producing risk-adjusted performance profiles. The simple example above ignores the need for risk adjustment by targeting patients with identical clinical conditions. In most situations, different providers see different mixes of patients, so risk adjustment is essential.

Linear regression modeling is the most commonly used method for risk adjusting continuous outcomes (see Chapter 10). Thus, we might build a model as follows:

$$PRED_i = \hat{a} + \Sigma_j \hat{b}_j * X_{ij}$$

where $PRED_i$ is the expected outcome for patient i, who has characteristics X_{ij}, $j = 1, \ldots, J$ for the J predictors in the model. For patients treated by a specific provider, their expected outcome (E) equals the average of the $PRED_i$s.

In contrast, logistic regression, in which the log of the odds of the event is modeled as a linear function of the predictor variables, is generally used to predict dichotomous (yes/no) outcomes.[6] After fitting a logistic regression model, the predicted probability of death for the ith case ($PRED_i$) is calculated from the relationship:

$$\ln(\widehat{\text{odds}}_i) = \ln(PRED_i/(1 - PRED_i)) = \hat{a} + \Sigma_j \hat{b}_j X_{ij}$$

by solving for $PRED_i$:

$$PRED_i = e^{\ln(\widehat{\text{odds}}_i)}/1 + e^{\ln(\widehat{\text{odds}}_i)}$$

To determine the expected number of deaths in a group of n cases, we sum the $PRED_i$ terms; to determine the expected death rate, we divide this sum by n.

Comparing observed to expected outcomes is central to performance profiling (e.g., drawing inferences about the quality or efficiency of care). Various approaches have been used for comparing O and E. Neither ($O - E$) nor O/E is clearly superior. For example, suppose that hospital A treats cases expected to average $5,000 (i.e., average $PRED_i$), but the actual cost is $6,000. In contrast, hospital B treats cases that should cost $10,000 but actually average $11,500. Thus, both hospitals' costs are greater than expected, but how do they compare with each other? On an additive or difference scale ($O - E$), hospital B performs worse, as hospital A's excess is only $1,000 per case, compared to B's excess of $1,500. However, on a multiplicative or ratio scale (O/E), hospital A does worse, as its cases cost 20 percent more than expected, compared to only 15 percent more for B. Theory offers no insight into which hospital to prefer.

Consider another example. Which is worse: 2 percent complications when only 1 percent was expected (double the rate), or 50 percent complications when only 40 percent was expected (ten excess problems per 100 but only a 25 percent higher complication rate)? This question has no simple answer. Analysts can use their data to explore which model is more realistic—an additive model (where adding the same amount to each case represents the provider effect) or a multiplicative model (where provider-associated increases are proportionate to the expected outcome). Even when multiplicative models are chosen, observers typically still want to know how observed results compare additively to expected results, such as how many extra dollars a provider costs or how many extra complications have occurred.

Ratios, such as O/E, are centered at 1 but range from 0 to infinity. To put comparable distances between ratios below 1 and those above 1, analysts sometimes display $\log(O/E)$ values rather than O/E values (Roos, Wennberg, and McPherson 1988). On a graph where O values are on the x-axis and $\log(O/E)$ values are on the y-axis, a "broken" y-axis can be used to indicate the gap between the smallest $\log(O/E)$ associated with a positive observed

(O) and negative infinity (the value of the logarithm function at zero). On an untransformed scale, substantial differences among O/E values less than 1 are hard to see and thus may appear unimportant. In contrast, on a log scale, the distances between points representing O/E values of 0.25, 0.50, 1.00, 2.00, and 4.00 are equally spaced because each value doubles the one below it.

A drawback of the ratio O/E is that when E is small, its value changes dramatically with small changes in O. For example, if we observe 30 cases, each with a 1 percent risk of complications, the expected number of complications is 0.3. If 0, 1, or 2 complications are observed, O/E is 0, 3.3, or 6.7, respectively. A good guideline is to avoid examining such ratios when the expected number of events is less than 1.0. Some researchers advise against O-to-E comparisons unless E is at least 5.

Fortunately, when comparing O to E, findings as extreme as our examples above are unlikely. If expected costs at two hospitals are $5,000 versus $10,000, or if expected complications rates are 1 percent versus 40 percent, these hospitals should probably not be compared—their patient populations or other characteristics differ too much. When distributions of expected outcomes are roughly similar across hospitals, difference and ratio measures of performance will produce reasonably similar results. Examining expected outcomes across providers is therefore important to ensure that patients' risks do not differ radically across providers (see below). Reviewing common descriptive statistics (e.g., mean and median, SD, percentage cutoffs of the distribution, boxplots) is a useful first cut at comparing expected outcomes across providers.

In our prior work, we examined the extent to which severity explained differences in hospital LOS for pneumonia patients (Iezzoni et al. 1996c). To illustrate the relationship between the distribution of predicted LOS and severity, we examined these distributions for six hospitals (number of cases per hospital ranged from 73 to 316): the three hospitals with the highest and three with the lowest predicted average LOSs (which corresponds to the hospitals with the lowest and highest risk-adjusted severity). Figure 12.1 shows side-by-side boxplots of predicted LOS values at these six hospitals, using DS Relative Resource Scale as the severity measure. (We removed outliers using Medicare's definition: cases more than three SDs above the mean on a log scale; see Chapter 10.)

In Figure 12.1, the box shows the range encompassing the middle 50 percent of cases. Thus, 25 percent of cases have values below the bottom edge of the box, and 25 percent have values above the upper edge. The horizontal line within the box is the median. The length of the box is the interquartile range (IQR), sometimes called the "H-spread." The top of the box plus 1.5 * IQR and the bottom of the box minus 1.5 * IQR define the inner fences; the top of the box plus 3 * IQR and the bottom of the box minus 3 * IQR define the outer fences. The ends of the lines extending above and below the box indicate the highest and lowest values within the inner fences; circles indicate

individual values between the inner and outer fences. (Different computerized statistical packages use different symbols, but the boxplot concept is similar.) The boxplots show that 75 percent of patients in the three hospitals with the least severely ill patients were expected to have an LOS below eight days, whereas 75 percent of cases at the hospitals with the most severely ill patients were expected to have an LOS above eight days.

Failure of O-to-E Comparisons to Adjust Fully for Risks

When examining death rates, epidemiologists often use standardized mortality ratios (SMRs). SMRs are *O/E* ratios, where the *E* values are calculated using indirect standardization. To illustrate the need for standardization (the epidemiologist's term for risk adjustment), consider a hypothetical situation involving two types of patients: low-risk, with a 1 percent mortality rate, and high-risk, with a 5 percent mortality rate (Table 12.4). Suppose further that half of all patients in a large population are low- and high-risk, yielding an overall mortality rate of 3 percent. Now consider hospital A, which treats 1,000 patients, 800 at low risk and 200 at high risk. Hospital A's experience with its low-risk patients is similar to the overall experience—a 1 percent mortality rate (8 deaths among the 800 patients). However, hospital A does poorly with high-risk patients; it has a 10 percent mortality rate, double the population average, leading to 20 deaths among the 200 high-risk patients. Despite this, because of its favorable case mix, hospital A's mortality rate is 2.8 percent (28/1,000), somewhat better than the 3 percent population average.

Indirect standardization determines a hospital's expected number of deaths by applying stratum-specific rates determined from all patients to the number of cases in each stratum in the hospital. In this case, a stratum is a risk category. Based on the overall data, we expect 8 deaths among the 800 low-risk patients (with a 1 percent mortality rate) and 10 deaths among the 200 high-risk patients (with a 5 percent mortality rate), for an expected rate of 1.8 percent.[7] The standardized mortality ratio for hospital A is 1.56 (28/18), since it has 56 percent more deaths than expected based on its patient mix.

One can report this discrepancy in other ways. Some prefer to express the hospital's performance on the same scale as the population average, giving a "risk-adjusted average." This is achieved by multiplying the SMR by the population average rate (e.g., $1.56 * 3 = 4.68$ percent). Another choice is to report the difference between the observed rate (2.8 percent) and the expected rate (1.8 percent); thus, hospital A has 1 percent more deaths than expected. All of these summary measures agree on the main point: After adjusting for its patient mix, hospital A has more deaths than expected.

Indirect standardization and its generalization via multivariable risk-adjustment modeling are powerful tools for making fairer comparisons among providers with different types of patients. Nevertheless, comparing outcomes across providers is complicated when patient mix both strongly affects the out-

TABLE 12.4
Hypothetical Hospitals with Different Patient Mixes and Death Rates

	All Patients in Population		Hospital A		Hospital B		Hospital C	
	Patient Mix (%)	Death (%)	*n*	Death (%)	*n*	Death (%)	*n*	Death (%)
Risk Category								
Low	50	1	800	1	200	1	800	1.25
High	50	5	200	10	800	10	200	12.50
Performance								
Observed death rate (*O*)		3	28/1,000 = 2.8%		82/1,000 = 8.2%		35/1,000 = 3.5%	
Standard mortality ratio (SMR) = *O*/*E*			28/18 = 1.56%		82/42 = 1.95%		35/18 = 1.94%	
Risk-adjusted mortality			3 * 1.56 = 4.68%		3 * 1.95 = 5.85%		3 * 1.94 = 5.82%	
Difference (*O* − *E*)			2.8 − 1.8 = 1%		8.2 − 4.2 = 4%		3.5 − 1.8 = 1.7%	

come and differs widely across providers. In the terminology of epidemiology, patient mix is a confounding factor when examining patient outcomes.

To illustrate, consider another institution, hospital B, with exactly the same mortality experience within each stratum as hospital A above, but with an unfavorable case mix. Hospital B treats 200 low-risk patients with 2 deaths and 800 high-risk patients with 80 deaths (see Table 12.4). Solely because of differences in patient mix, hospital B's unadjusted death rate is 8.2 percent, much higher than hospital A's 2.8 percent rate. When facilities differ widely in their patient mix, "raw" comparisons can mislead.

However, risk adjustment does not always do what we anticipate or hope that it does. For example, an indirect adjustment approach fails to make hospitals A and B look equally good. To perform indirect adjustment for hospital B, we first compute its expected number of deaths as 42 (0.01 * 200 + 0.05 * 800). Hospital B's SMR is thus 1.95, its risk-adjusted death rate is 5.85 percent, and its excess mortality rate is 4 percent (as opposed to 1.56 percent, 4.68 percent, and 1 percent, respectively, at hospital A). However reported, hospital B looks worse than A, although the same type of patient had the same outcome at either hospital. Results could be even more misleading. Imagine that hospital C is seriously deficient: It has the same favorable patient mix as hospital A but 25 percent higher death rates for both patient types (1.25 percent and 12.5 percent mortality, respectively, among low- and high-risk patients). Hospital C's SMR, risk-adjusted death rate, and excess mortality rate (1.94 percent, 5.82 percent, and 1.7 percent, respectively) look marginally better than hospital B's, although its performance is clearly worse.

Direct standardization, an alternative adjustment approach, produces results that feel more correct, but the method has conceptual and practical problems. In direct standardization, provider-specific rates are computed in

each risk stratum and applied to a "standard" population case mix, producing an estimate of what might be expected if the provider were to treat this standard patient mix. For example, suppose that the standard population has 50 percent low- and 50 percent high-risk patients. Under this assumption, hospitals A and B have stratum-specific death rates that are estimated to yield 5.5 percent mortality in the standard population ($0.5 * 0.01 + 0.5 * 0.10$), as compared to hospital C's estimated 6.9 percent rate ($0.5 * 0.0125 + 0.5 * 0.125$).

In epidemiological studies, the strata are generally large, such as populations in different states broken into five-year age categories. Relatively reliable estimates of stratum-specific rates are possible using such large populations. However, in profiling individual providers for patients stratified by disease or other risk factors, stratum-specific rates are generally based on too few cases to provide reliable estimates. Furthermore, questions arise about whether a provider should be judged harshly for ostensibly doing poorly with types of patients that it rarely sees. For example, suppose hospital D treats 1 high-risk patient who dies and 999 low-risk patients, of whom only 5 die. Although its death rate is only half as large as the 1 percent expected rate for nearly all of its 1,000 patients, its projected death rate for the standard population is more than 50 percent ($0.5 * 0.005 + 0.5 * 1.00$). Thus, as this example demonstrates, which of several providers looks best can change depending on the patient mix of the standard population. Direct standardization is rarely used to profile physicians or hospitals.

Most performance profiles use more complex multivariable models to determine expected values. However, the fundamental approach is identical: Each provider's observed outcome is compared to expected outcomes based on the risk characteristics of its patients and the model-specified relationship between these characteristics and the outcome of interest. When providers treat very different populations, risk adjustment therefore cannot answer definitively "which is better." In reality, particular providers may do better with certain types of patients and worse with others. Thus, examining the data in multiple ways becomes particularly important, for example, examining provider performance within high- and low-risk strata of patients. In a rational world, providers would concentrate on their most successful types of cases, and performance profiles would help steer patients to providers who do well with similar kinds of patients.

Random Variation in Comparing O-to-E Outcomes

As discussed, standard errors capture the effect of random variation on the reliability of estimates from data. When each of n observations is an independent observation from a common distribution with mean μ and SD σ, we estimate μ by $\overline{Y}$ and σ by:

$$\hat{\sigma} = s = \sqrt{\Sigma_i (Y_i - \overline{Y})^2/(n-1)} \qquad (1)$$

The values of $\overline{Y}$ and s remain relatively constant as n increases, each becoming an increasingly accurate estimator of μ and σ, respectively. In contrast, the statistic that measures the accuracy of $\overline{Y}$ as an estimator of μ becomes smaller as n increases:

$$\mathrm{SE}(\overline{Y}) = s/\sqrt{n}$$

Properly estimating standard errors for predictions of providers' expected outcomes requires care. Consider predictions of a continuous outcome from a multivariable linear regression model. Most computerized statistical regression packages estimate the SE associated with an observed outcome. However, there are two SE values: the SE for the expected value of an observed outcome (i.e., for the mean of many patients similar to that one for whom the prediction is made), and the SE for that individual patient. For provider profiling, the latter SE is more relevant. It is larger than the first because it reflects not only uncertainty about estimates of parameters in the model but also uncertainty associated with the outcome of an individual observation given its expected value.

If s_i is the standard error for the ith observation, the SE for the average of a group of n cases is:

$$\frac{\sqrt{\Sigma_i s_i^2}}{n}.$$

An approximate 95 percent CI for the average outcome of n patients at hospital A is:

$$\overline{Y}_A \pm 2 * \frac{\sqrt{\Sigma_i s_i^2}}{n}.$$

The distributions of continuous outcomes like costs and LOSs usually have "long right tails," including some cases with extremely high values (see Chapter 10). Therefore, the logarithm (usually the natural logarithm) of these continuous values is often used in modeling because its distribution is more symmetrical than that of the untransformed data. In this situation, CIs can be computed on the log scale. However, achieving estimates on the original scale requires that the "point estimate" of the mean and the endpoints of the CI be retransformed by exponentiation (and adjusted for bias; see Chapter 10). Resulting CIs will not be centered at the estimated mean.

Consider a dichotomous outcome like death. The SD associated with an individual's predicted probability of death (corresponding to s_i for a continuous dependent variable) is:

$$\sqrt{\hat{P}_i\,(1 - \hat{P}_i)}$$

and the SD of the mean death rate of n persons, used to estimate the "real" death rate P that we cannot observe, is:

$$\frac{\sqrt{\Sigma_i \hat{P}_i (1 - \hat{P}_i)}}{n}.$$

When each provider's expected outcome (E) comes from a model fit to many cases, calculations of 95 percent CIs for O/E ratios can treat E as a constant. Then, one can calculate the CI around the observed number of deaths as:

$$O \pm 2 * \sqrt{\Sigma_i \hat{P}_i (1 - \hat{P}_i)}$$

and divide the resulting lower-, midpoint-, and upper-interval values by E. Multiplying each end of the CI by the area-wide rate yields a CI for risk-adjusted outcomes. Hosmer and Lemeshow (1995) found this approach was reasonable based on simulation studies, including a situation in which the observed cases were excluded from the data set used to build the model generating the expected findings.

Presenting Comparisons of O-to-E Outcomes

As noted throughout this book, risk-adjusted outcomes information—including performance profiles—is increasingly used by nontechnical audiences for a variety of purposes. Therefore, results must be presented in a clear-cut, easy-to-understand fashion. Because few users comprehend the methodological underpinnings of the computations, some performance profiles aim for the simplest presentation, even though it obscures important issues. For example, the 1996 profile of health plan performance produced by the Massachusetts Healthcare Purchaser Group initially arrayed ratings on a scale of one to five stars, establishing cutpoints to determine the numbers of stars. Apparently some health plans objected, noting that a single star, for the lowest-rated plan, sent a disproportionately negative message. The final version of the rating used only the center part of the five-star scale, with the lowest-rated plan having two stars and the highest having four. The *Boston Globe* published this rating on the front page of their business section, with the stars printed in bright red (Pham 1996a, C1). The numerical rankings appeared to the right of the stars, in small black print on a gray background. The lowest three-star plan had 89.8 percent overall satisfaction, while the sole two-star plan, Massachusetts Blue Cross and Blue Shield, had 87.3 percent, hardly a striking difference. Blue Cross withdrew from the voluntary rating program, noting that their low rating by the star method obscured the fact that their performance was numerically only slightly below their competitors' (Pham 1996b, C3).

However, the American public frequently sees results of opinion surveys presented alongside their "margin of error," especially around election time. People should therefore not have difficulty understanding that comparisons of observed-to-expected outcomes are also uncertain. How should this uncertainty be portrayed on the printed page? Producers of performance profiles use

two common strategies to depict differences between O and E outcomes in a way that captures the effect of random variation: showing the observed value in relationship to a "prediction interval" of the form $\mu \pm 2 * \text{SE}$ or measuring the difference between observed and expected values in units of SE.

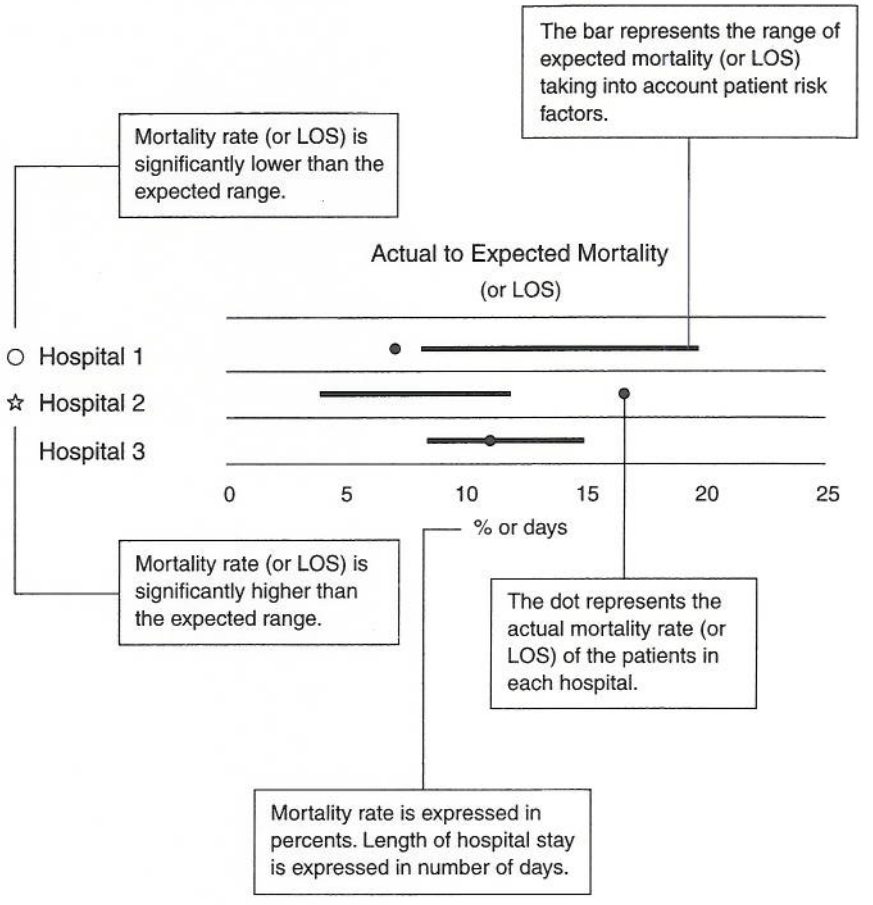
A report card on heart attack outcomes produced by the Pennsylvania Health Care Cost Containment Council (PHC4 1996) illustrates the first approach. PHC4 developed separate models for "direct admits" (patients receiving initial care for a heart attack) and for "transfer-ins." Figure 12.2, reproduced from the PHC4 report, illustrates a prediction interval generated from a multivariable model that adjusts for risk factors and a particular hospital's rate relative to the interval. Figure 12.3, also from the PHC4 report, shows an example of the results. The prediction intervals were wide, many spanning a range of 10 percent (e.g., from 5 percent to 15 percent). The interval for Aliquippa Hospital is typical; based on 87 cases, the interval ranged from about 3 percent to 13 percent. Butler Hospital, with 259 cases, had a narrower interval (from about 6 percent to 12 percent), whereas Corry Memorial Hospital's interval, based on only 46 cases, went from about 2 percent to 17 percent. Hospitals with higher or lower observed than expected rates are obvious, flagged by a small symbol to the left of the hospital.

For all hospitals depicted on the same page, readers can compare the relative widths of the prediction intervals, which are primarily a function of the number of cases treated. Hospitals with wider intervals treat fewer cases. Despite cluttering the presentation, showing the number of cases treated at each hospital would have been useful here, although detailed tables later in the report list the number of cases, percentage transferred out, and actual values with 95 percent CIs for mortality and LOS (PHC4 1996, 23–25).

In Figure 12.3, 4 of the 39 hospitals had rates outside the 95 percent prediction intervals, one with a higher rate than expected. The display invites the conclusion that the three below-interval hospitals had particularly high-quality care and the one above-interval hospital had low-quality care; however, 5 percent of ordinary hospitals fall outside the 95 percent prediction interval because of random chance alone. Thus, among 39 similar hospitals, about two would be spuriously flagged. Of the four that fell outside the interval, readers cannot know which, if any, are quality outliers. The traditional approach for adjusting for this "multiple comparisons" problem essentially expands the width of the prediction interval. This decreases the power of tests to identify hospitals that really do differ from expected and has been rarely used in performance profiling. As discussed later, hierarchical models handle the multiple comparisons problem better.

The PHC4 heart attack report showed the same findings for physician practices within hospitals, provided they treated more than 30 cases. As noted earlier, if a continuous variable does not have a widely skewed distribution, the mean of a sample of 30 cases is approximately normally distributed, probably the rationale for choosing a minimum sample size of 30. However, in this

FIGURE 12.2
Instructions for Reading PHC4's Hospital Performance Reports



SOURCE: Pennsylvania Health Care Cost Containment Council (1996).

situation, PHC4 addressed a dichotomous outcome. With such an outcome, the normal approximation is generally reasonable when the event rate times the number of cases is at least five. Using this rule, 30 cases necessitate an event rate of more than 15 percent. This is higher than what occurred (Localio et al. 1997).

In both the mortality and LOS analyses, the PHC4 excluded patients from hospitals closing since 1993; who left against medical advice; under 30 and over 99 years of age; from hospitals treating fewer than 30 cases; involved in two or more transfers; and who were "clinically complex," with a preexisting or coexisting clinical condition not related to heart attack treatment and not included in the risk model. The LOS analyses also excluded patients who died, patients transferred out (who had "truncated" LOSs), and patients with "atypical" LOSs (more than 40 days or those discharged on the same day they were admitted). While PHC4 exempted hospitals with fewer than 30 cases from the mortality analysis, they included all hospitals in the LOS analysis.

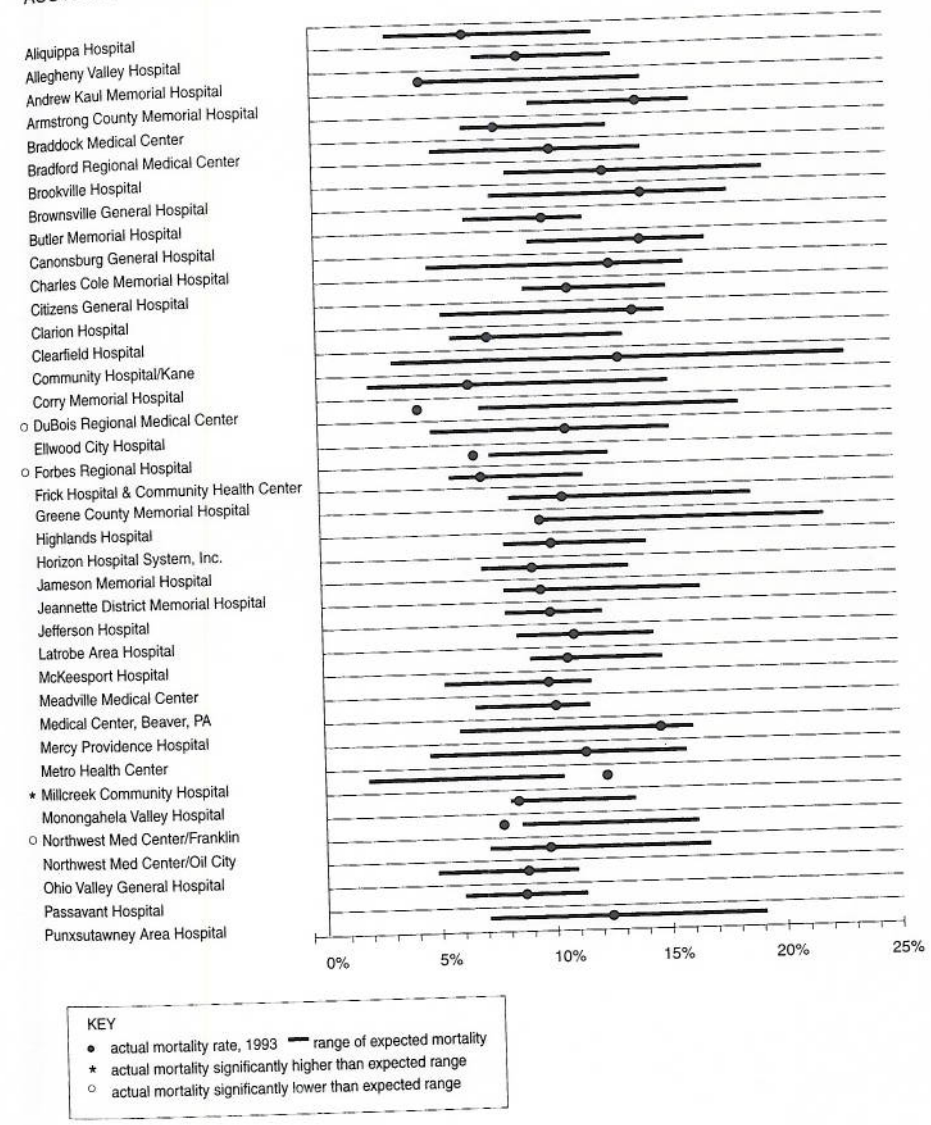


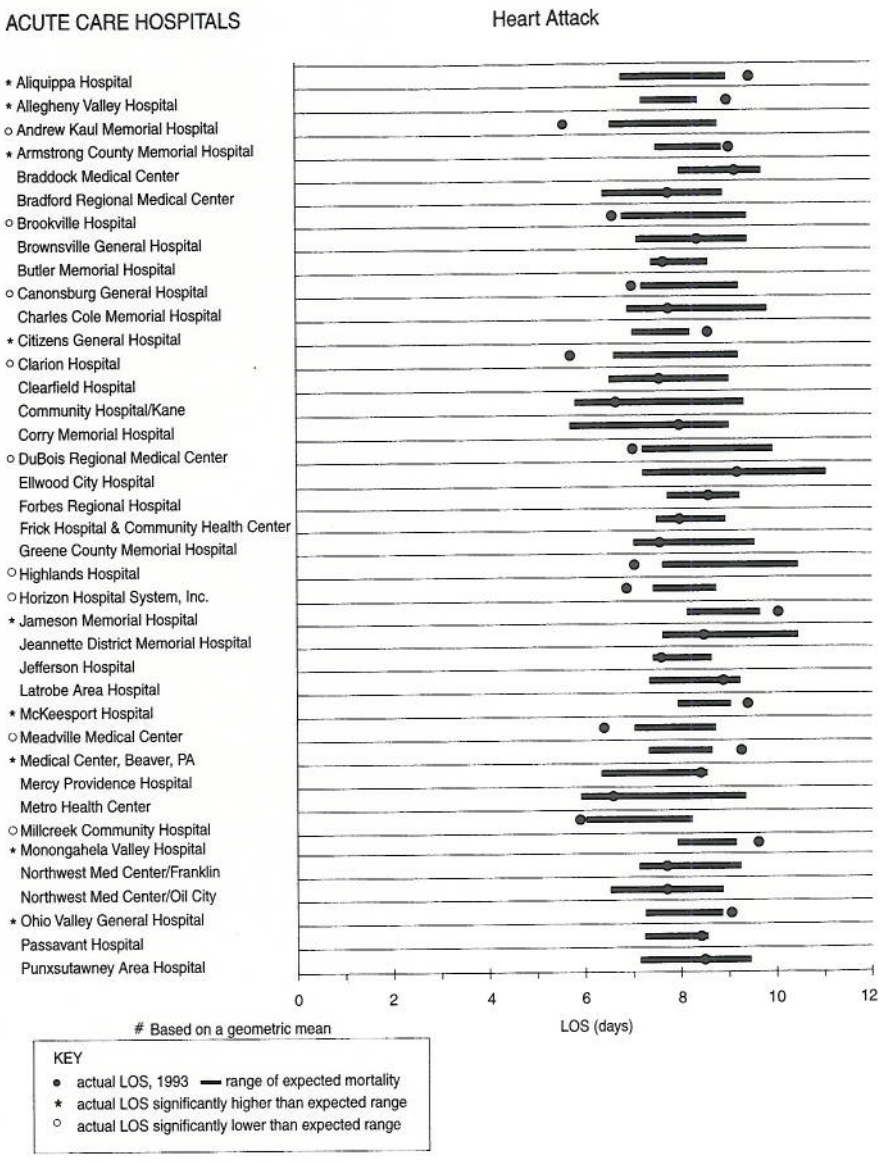
FIGURE 12.3 Mortality Rates and Prediction Intervals for a Sample of Pennsylvania Hospitals

SOURCE: Pennsylvania Health Care Cost Containment Council (1996).

The PHC4 LOS analyses used a log transform: ln(LOS) was the dependent variable. They calculated upper and lower endpoints for CIs in the log scale, then retransformed them by exponentiation. As illustrated in Figure 12.4, presentation of LOS data paralleled that for mortality. More hospitals fell outside the prediction intervals for LOS than for mortality.

Obviously, the number of cases treated is crucial in interpreting such data. For example, if a provider's expected death rate is 10 percent, an observed rate of 15 percent based on 400 cases is more worrisome than either an observed rate of 15 percent based on 100 cases or an observed rate of 20

FIGURE 12.4 Average LOS and Prediction Intervals for a Sample of Pennsylvania Hospitals



SOURCE: Pennsylvania Health Care Cost Containment Council (1996).

percent based on 10 cases. Figures 12.3 and 12.4 appropriately remind readers to pay less attention to deviations based on fewer cases.

Standardizing is a common statistical technique for converting a deviation (i.e., an $O - E$) into a measure that suggests whether the deviation is statistically meaningful. We consider:

$$z = (O - E)/\text{SE}$$

where SE is calculated as described above. If the observed rate pertains to a process whose expected value really is E, and if n is sufficiently large, this

quantity has approximately a standard normal distribution. This is called a "z-score," since z is used in statistics to denote the standard normal random variable. The standard normal is centered at 0; 68 percent of z-scores fall in the interval from –1 to +1; and slightly more than 95 percent are in the interval from –2 to +2. Widely available standard normal tables (or computer programs) are used to convert z-scores into p-values.

The p-value measures how likely it is for observed and expected rates to differ at least as much as they do, assuming that the observations reflect the hypothesized model. For example, a standard normal variate falls outside the interval –1 to +1 only 32 percent of the time. Thus, a z-score of either +1 or –1 (i.e., O and E differ by one SD) has $p = 0.32$. A z-score greater than 1.96 or less than –1.96 has a p-value smaller than 0.05. If a provider's $O - E$ value leads to $z = 1.96$ ($p = 0.05$), its true rate may nonetheless be E. However, random deviations this large occur only one time in 20.

We used z-scores in study of hospital-based severity measures. Table 12.5 shows z-scores at five hospitals when we used different severity measures to determine expected rates of death for pneumonia patients (Iezzoni et al. 1996a). We selected these five hospitals for illustrative purposes from the 30 out of 105 hospitals in the study at which observed mortality rates differed significantly from expected ($p < 0.05$) when judged by one or more, but not all 14, severity methods analyzed. For example, at hospital B, observed mortality was significantly lower than expected when using DS's probability of mortality model ($z = -3.07$, $p < 0.01$), APR-DRG ($z = -2.30$, $p = 0.02$), or PMC-RIS ($z = -2.16$, $p = 0.03$). In contrast, observed rates were less than two SEs from expected, consistent with the null hypothesis of no difference, when using the original version of MedisGroups ($z = -1.33$, $p = 0.18$), physiology score 1 ($z = -1.53$, $p = 0.13$), or R-DRG ($z = -0.84$, $p = 0.40$), as well as other severity measures. Thus, whether hospital B was identified as a particularly high quality hospital, and perhaps used to benchmark performance at other institutions, depended on which severity measure was used for risk adjustment.

California's hospital report card initiative used a similar approach to portray outlier hospitals (Wilson, Smoley, and Werdegar 1996). However, rather than use the normal approximation to convert a z-score into a p-value, an "exact" p-value was calculated as described in Luft and Brown (1993). Figure 12.5, taken from the California report, shows how they reported their results.

Some critics complain that z-scores and p-values are not intuitive; many consumers of report cards would rather receive their information in such familiar terms as rates of excess problems or risk-adjusted problem rates. A deeper criticism is that any single-number summary (or point estimate) is likely to convey more precision than is justified, especially when ranking providers. Several strategies might combat this "tyranny of spurious precision." One involves using categorical reporting, as in Figure 12.5. This solves the problem of believing that one provider is better than another because 3.2 is bigger

TABLE 12.5
Examples of Relative Mortality Rate Performance from Five Hospitals: Pneumonia Patients

Hospital Performance Measure and Severity Method	A	B	C	D	E
No. of cases	200	317	88	267	132
No. (%) died	17 (8.5)	32 (10.1)	10 (11.4)	36 (13.5)	25 (18.9)
z-score (decile rank) from unadjusted model[a]	−0.53 (4)	0.29 (7)	0.56 (8)	2.14 (10)	3.63 (10)
	z-score (decile rank) from severity-adjusted model[b]				
MedisGroups	−2.30 (1)	−1.33 (2)	1.56 (9)	2.70 (10)	1.99 (10)
Original version	−2.73 (1)	−1.73 (1)	2.03 (10)	1.33 (9)	1.17 (9)
Empirical version	−2.25 (1)	−1.53 (1)	1.64 (9)	2.24 (10)	2.79 (10)
Physiology score 1	−3.12 (1)	−1.84 (1)	1.49 (9)	1.95 (10)	2.02 (10)
Physiology score 2	−1.74 (1)	−1.23 (3)	1.95 (9)	2.29 (10)	3.07 (10)
Body Systems Count	−1.28 (2)	−1.13 (3)	1.32 (9)	2.11 (10)	3.16 (10)
Comorbidity Index					
DS	−2.51 (1)	−3.07 (1)	1.51 (9)	3.51 (10)	2.14 (9)
Mortality probability	−2.05 (1)	−0.95 (3)	1.87 (9)	2.12 (10)	2.88 (10)
Stage	−1.15 (2)	−1.66 (1)	1.45 (8)	2.05 (10)	2.57 (10)
Comorbidities	−1.99 (1)	−1.88 (1)	2.23 (10)	1.04 (9)	2.78 (10)
PMCs: severity score	−1.54 (1)	−1.97 (1)	1.99 (9)	2.17 (10)	3.20 (10)
AIM	−2.25 (1)	−2.30 (1)	2.32 (10)	1.73 (9)	2.50 (10)
APR-DRGs	−2.05 (1)	−2.16 (1)	2.60 (10)	1.48 (9)	3.41 (10)
PMC-RIS	−2.08 (1)	−0.84 (3)	0.60 (8)	3.04 (10)	1.79 (10)
R-DRGs					

SOURCE: Iezzoni et al. (1996a).
[a] Unadjusted model assumed 0.096 probability of death for all patients.
[b] Severity-adjusted model included age-sex, DRG, and severity score.

than 3.1, for example, at the cost of introducing harmful "edge effects" in which two nearly identical hospitals appear different because only one "made the cut" into a better category (as illustrated in the *Boston Globe* example at the start of the section). Another strategy is to report numerical performance measures using fewer decimal places (e.g., 3.1 rather than 3.14159). Displaying confidence (or acceptance) intervals is also helpful. Finally, performance reports should resist the urge to list providers in rank order from "best to worst."

Pictures often convey messages more powerfully than words or numerical tables. The graphical displays shown in Figures 12.3 and 12.4, for example, provide point estimates and prediction intervals for each hospital in a way that facilitates and encourages comparison. One important feature of such displays involves the order for listing data from different hospitals. The Pennsylvania report used alphabetical order, making it easier to locate

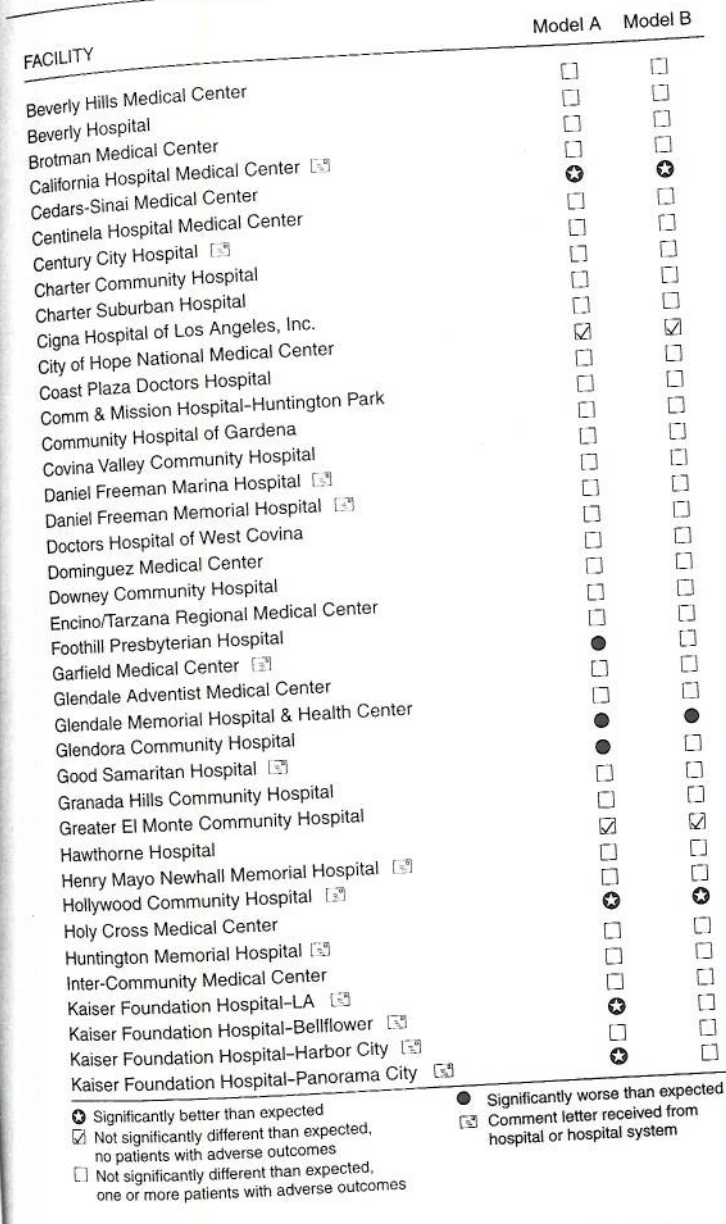


FIGURE 12.5 Portraying "Outlier Status" for a Sample of California Hospitals

SOURCE: Wilson, Smoley, and Werdegar (1996).

information for a particular hospital and harder to find the supposedly best and worst performers.

Displaying hospitals from highest to lowest performance draws attention to ranking, an unwise choice given the unreliability of rank determinations (Goldstein and Spiegelhalter 1996). Furthermore, when observed rates, *O/E* ratios, or risk-adjusted rates are used to establish a rank ordering, the most extreme rates are usually those based on few cases (such as 0 percent problems based on zero of ten cases). These extremes most likely reflect randomness

and will likely change in subsequent periods. Reordering the same data may prompt new insights. For example, displaying hospitals by important characteristics (e.g., ownership, payer mix, teaching intensity) encourages comparisons among similar facilities. Such displays also highlight differences by type of hospital.

The art and science of good visual displays has advanced rapidly in recent years (Tufte 1983). Also, software for producing graphics is increasingly available, such as the many powerful display formats that have been implemented in the S or S+ computer language (Cleveland 1993).

Figure 12.6 shows how one might compare several providers (e.g., hospitals A, B, and C) on their performance with each of several kinds of patients (e.g., low, middle, and high risk) by portraying results separately within patient categories (Teres and Lemeshow 1993). Observing a multidimensional "signature" for a hospital might highlight areas that require explanation and reveal potential strategies for improvement (e.g., hospital A may do well with low-risk patients and poorly with others). Such an approach, however, requires enough patients to estimate rates reliably within each hospital or risk cell. Almond and colleagues (2000) suggest useful graphical displays for comparing a particular provider to other providers in the group.

Bayesian Models

Standard approaches for provider profiling present several problems. The first is how the "true" mean value of an outcome is estimated for each provider (e.g., μ_A for provider A). Traditionally, this is calculated separately for each provider as the average outcome of patients treated by that provider ($\overline{Y}_A$). However, the resulting set of provider estimates is often not as close as it could be to the true means and not the best predictor of what will happen in the future. Typically, $\overline{Y}_A$s are too extreme, the highest ones being higher than the associated true μ_As and the lowest $\overline{Y}_A$s being too low. When provider-specific averages are based on small numbers of patients, large over- and underestimates are especially likely.

In addition, traditional estimates of SE values (described above) may understate the amount of variability that is present, leading to CIs that are too narrow. Understating true variation causes more normal providers than expected to be flagged as outliers. One reason is that traditional methods recognize only one source of variation in the data—random variation of patients within a provider. However, variation across providers also exists. Standard errors are also often underestimated because patients treated by a particular provider may fall into groups such that patients in one group are more similar to each other (for whatever reason) than patients in another group. In other words, patients may not be independent observations but are "nested" or "clustered," often by some organizational hierarchy: For example, patients are clustered by their treating physicians; similarly, CABG surgeons are nested

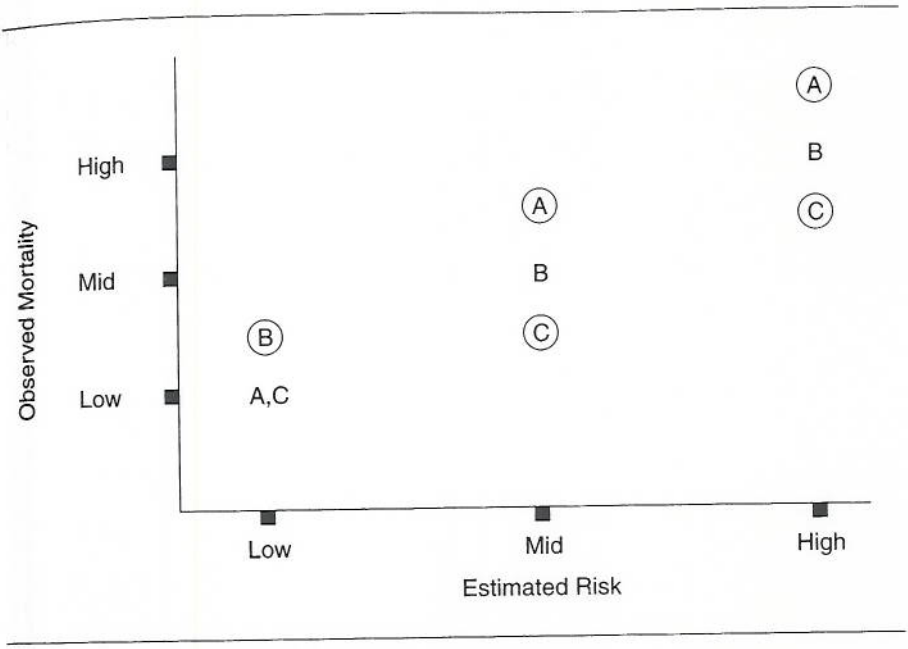


FIGURE 12.6 Portraying Outcomes by Risk Strata

SOURCE: Teres and Lemeshow (1993).

within hospitals. When analyzing units within which clustering occurs, effective sample sizes (in terms of the amount of information provided) are less than actual sample sizes. Approaches that do not adjust for clustering underestimate SEs (Greenfield et al. 2002). Bayesian hierarchical models (also called multilevel or random coefficient models) provide a comprehensive approach for dealing with such problems.[8]

The Bayesian Approach

Hierarchical models fit within a more general Bayesian perspective that views newly observed data within the context of prior knowledge. For example, suppose in ten coin tosses we observe one head. We know from our understanding of coin tosses that P (the true probability that a coin comes up heads) is quite close to 0.5. Therefore, our observed rate of 0.1 heads raises some question about whether the coin is truly fair.[9] However, if the coin looks normal, it still seems more reasonable to suppose that the P is more like 0.5 than 0.1. If we toss the coin more and continue to generate fewer than 50 percent heads, we grow more suspicious that P is less than 0.5. If the coin is biased, what is its true P? As the number of coin tosses increases, the observed proportion of heads increasingly becomes a more credible estimate of P.

Similar reasoning is useful for evaluating providers such as hospitals, where, without data, we assume that hospital A is ordinary and has outcomes like those at other hospitals. (Technically, this is called the exchangeability assumption.) If we receive a little evidence about hospital A's performance, we adjust this "prior" estimate slightly. As evidence accumulates, however, we place increasing weight on the new data and less weight on our prior belief.

At some point, the data may be enough to convince us that hospital A is of truly lower (or higher) quality or efficiency than other hospitals.

This is the sense in which Bayesian analyses interpret new data within the context of prior beliefs. To give another example, if we are uncertain about the safety of an operation but we observe one complication in ten operations, we may accept $\hat{P} = 0.1$ as our best guess for the true complication rate P. However, if we observed no complications in ten operations, we may feel uncomfortable with $\hat{P} = 0$, since we know that all surgery presents risks. Implicit in this is our belief that this operation is in some ways like other surgeries where complication rates around 10 percent are typically seen, but a complication rate of 0 percent is not.

Classical statistical methods capture the level of uncertainty in estimating P by putting confidence intervals on $\hat{P}$ (as described above). This approach has two major limitations: First, computations of $\hat{P}$ and a CI for P rely only on current observations; and second, the approach leads to "all-or-nothing" estimates. That is, if the difference between observed and expected values is statistically significant (e.g., because a 95 percent CI does not include the expected value or the p-value is < 0.05), we accept the observed mean as the best guess for the true mean. Otherwise, we continue to believe that the true mean equals the previously held expected value. Using traditional hypothesis testing, our ten coin tosses lead us to estimate P as 10 percent if we observe one head (i.e., after rejecting the null hypothesis that P is 0.5). However, if we observe two to eight heads, we would conclude that the coin is probably fair, and our best guess for P would be 0.5.

A Bayesian framework uses prior knowledge about a situation to produce estimates for the true mean that lie somewhere between the observed average and the expected mean based on prior knowledge. The resulting estimate is closer to the expected mean when the observed mean is based on few data and prior knowledge (e.g., about how fair coins behave) is strong. The estimate is closer to the observed mean to the extent that either outside knowledge is less certain (e.g., surgery with an unknown complication rate) or when more data are available (e.g., when the observed mean derives from 1,000 operations rather than just 10).

Historically, the Bayesian dependence on "prior knowledge" has generated considerable controversy. Two people examining identical data might reach different conclusions because of personal differences in their prior knowledge and assumptions. Nevertheless, Bayesian analysis has entered mainstream statistics, partially because powerful modern computers have overcome previously intractable computational problems associated with the approach. Furthermore, analysts can use a Bayesian framework without depending strongly on prior knowledge. One way is through an approach called "empirical Bayes"—using the data both as the basis for prior knowledge and to adjust this knowledge. Another strategy is to assume very vague prior knowledge, captured by placing noninformative prior probability distributions on

unknown parameters (e.g., by conjecturing that complication rates associated with a new surgery are uniformly distributed between 0.1 percent and 80.0 percent). When prior knowledge is vague, the data primarily drive the Bayesian estimates.

Empirical Bayes Analysis

Casella (1985, 83) attributes the basis of "modern" empirical Bayes analysis to work by Efron and Morris (1972, 1973, 1975). As they discuss in an excellent nontechnical paper (Efron and Morris 1977), parametric empirical Bayes analysis derives from a theorem initially proven by Stein (1955) that challenged the fundamental assumption of traditional estimation theory, that the average of the observed data is usually the best estimate of the mean of the population from which the data were drawn. However, Stein proved that there are better ways to estimate jointly the (true) means of three or more normal populations than by using the three averages computed separately from samples from each population. The thinking inspired by Stein's theorem evolved into better ways to estimate means for several populations simultaneously.[10]

Empirical Bayes estimation extracts information from the current data set to function as the prior knowledge required for Bayesian estimates. For instance, in a particular database of 10,000 AMI patients treated at 100 hospitals, 13 percent die in hospital. This 13 percent serves as prior knowledge and provides a context for interpreting hospital A's experience—10 deaths out of 100 patients. The empirical Bayes estimate for the true death rate at hospital A will lie between 10 percent (the observed value) and 13 percent (our prior knowledge). Exactly where depends on the relative size of random variation in the observed death rate (10 percent) and the amount by which true hospital death rates differ from their mean of 13 percent.

To illustrate the empirical Bayes approach, consider comparing costs at four hospitals. A typical classical analysis considers two alternatives: (1) to accept the null hypothesis, in which case the true mean cost at each hospital is estimated as the common mean from the pooled sample of all patients, or (2) to reject the null hypothesis, in which case the mean of each hospital is estimated as the average of patients in that hospital. The empirical Bayes estimator represents a compromise, estimating each hospital's mean by giving weight to both the common mean and the mean at each hospital. Thus, the empirical Bayes estimate of the average cost in each hospital "shrinks" the hospital-specific cost toward the overall average.

Empirical Bayes estimates explicitly recognize two sources of variation in the data: (1) random variation within each unit examined (e.g., within each hospital, variation of individual patients' observed costs from the hospital's true average cost, measured by σ_A for hospital A) and (2) variation across hospitals in their true average costs (i.e., variation in the μ_i values for $i = 1$ to N, the number of hospitals). In making empirical Bayes estimates, the weight given to the observed mean in each unit is a function of these two sources

of variation, measured by the variance (which equals the standard deviation squared):

$$\text{weight} = \text{variance across units}/(\text{variance across units} + \text{variance within units})$$

As variation within units (e.g., hospitals) increases, unit-specific averages receive less weight (i.e., estimates are shrunk closer to the overall average). As noted earlier, variation in the average is $s/\sqrt{n}$, making within-unit variation larger for smaller samples. Usually, the most extreme raw averages come from units with small sample sizes. Thus, their Bayes estimates are shrunk much closer to the overall mean, leaving units with less extreme raw averages based on larger samples with the most extreme Bayes estimates.

We used empirical Bayes techniques to profile small geographical areas based on hospitalization rates among people aged 65 and over in Massachusetts (Shwartz et al. 1994). Specifically, we examined so-called "relative hospitalization rates" (RHR) in each geographical area, defined as the observed number of hospitalizations minus the expected, expressed as a multiple of expected:

$$\text{RHR} = (O - E)/E$$

Thus, for example, RHRs of –0.5, 0.0, and +0.5 represent areas with 50 percent less than expected, as much as expected, and 50 percent more than expected hospitalizations, respectively. We determined expected numbers of hospitalizations using indirect standardization to adjust for differences in age and sex distributions of the population in each area. Consider the effect of empirical Bayes shrinkage on perceptions of hospitalizations for cardiac catheterization. The highest RHR for cardiac catheterization, 0.90, occurred in a very small area with only 4,955 residents over age 64. The second highest cardiac catheterization RHR, 0.84, came from a much larger area, with 40,390 residents over age 64. The empirical Bayes estimates for the two areas were 0.65 and 0.80, respectively. Because the first area had a small population, the empirical Bayes estimate gave less weight to its observed rate and more to the overall mean of 0. In other words, the rate estimated by empirical Bayes "shrank" much closer to the overall mean, from 0.90 to 0.65. Because the second area had a much larger population, far less shrinkage occurred. This illustrates how empirical Bayes techniques adjust point estimates to reflect the uncertainty associated with raw averages, helping to guard against drawing conclusions from extreme estimates based on a few cases.

Our study also found that the set of empirical Bayes estimates of hospitalization rates in small geographical areas generally matched the set of area-specific rates for the following year better than did the raw averages (Shwartz et al. 1994). For 62 of the 68 conditions studied, empirical Bayes estimates yielded smaller weighted average errors (weighting by the size of the areas) when used to predict next year's hospitalization rates.

It is important to note that a shrunken (Bayes) estimate may be worse than the average of the data for an individual unit. Across all units, however, shrunken estimates produce lower overall errors.

Hierarchical Bayesian Models

Hierarchical models generalize the idea of shrinkage and provide a comprehensive framework for explicitly incorporating variation at different levels of analysis (Bryk and Raudenbush 1992). The "hierarchy" derives from nesting, which arises when the data are not generated independently, but in groups. For example, patients are nested within provider (a group of patients treated by the same physician); in turn, providers may be nested within practice groups (e.g., physicians who work at the same hospital) and hospitals may be nested within types (e.g., teaching versus nonteaching). At each level of the hierarchy, the relevant independent variables and their influence may differ. Explicit modeling of the hierarchical structure recognizes that nested observations may be correlated and that different sources of variation can occur at each level.

Consider the cost for treating patients with a particular disease; to simplify, assume patients are clinically similar across providers. For profiling provider costs, we could use the following simple (although slightly unrealistic) hierarchical model:

- Υ_{ij} has some distribution (e.g., normal truncated to be positive) with mean μ_j and variance σ^2 (stage I model)
- μ_j has some distribution (e.g., normal truncated to be positive) with mean λ and variance ν^2 (stage II model)

Υ_{ij} is the observed outcome for person i treated by provider j. Provider j's true expected outcome is μ_j. This model assumes that random variation of outcomes is identical for each provider. It is measured by σ^2. Providers' true expected outcomes differ; in this example, they follow a normal distribution with mean λ and variance ν^2. Hence, we can generate a data point as follows: (1) by randomly selecting a μ_j from a positively truncated normal distribution with mean λ and some variance ν^2 and (2) by randomly selecting a Υ_{ij} from a log normal distribution with mean μ_j and variance σ^2. Thus, the Υ_{ij} values have two sources of variation, one due to variation within provider (the σ^2) and another due to variation across providers (the ν^2). In a Bayesian hierarchical framework, the stage II parameters λ and ν^2 are called hyperparameters. These parameters are usually given vague or noninformative prior distributions (which implies vague prior knowledge); for example, the distributions are uniformly (or nearly uniformly) distributed over some appropriately wide range that incorporates any feasible values. As a result, the data primarily determine the estimates. Analysts can easily enrich this simple model by incorporating individual-level risk factors or a risk score in the stage I model and provider-level covariates (e.g., physician specialty, practice site) in the stage II model.

Hierarchical models have several key features (Thomas, Longford, and Rolfe 1994). They

- explicitly model differences among providers (over and above what is explained by differences in patient mix);
- view provider effects as "random variation," with the measure of spread, ν^2, estimated during model fitting;
- "shrink" the point estimate of a provider's outcome from the observed provider average toward a risk-adjusted expected value for the provider by an amount that depends on ν^2, σ^2, and the provider's sample size;
- produce wider intervals around point estimates that appropriately reflect the uncertainty arising from both individual variation of patients within providers and variation of providers; and
- provide a framework for comprehensively addressing the problem of multiple comparisons.

Gatsonis and colleagues (1995) offer a good nontechnical illustration of hierarchical modeling in examining variations across states in the use of coronary angiography for more than 218,000 elderly AMI patients. Patients were nested within state; states were nested within region. Within each state, the probability that a patient received angiography was modeled using logistic regression as a function of patient age, sex, race, and comorbidities. The researchers coded independent variables so that the intercept in that state was the log odds that a baseline case (a 65-year-old nonblack man with no comorbidities) received angiography. These were the stage I (or level 1) models. In stage II, they modeled the intercepts from the stage I models as a function of region and a measure of the availability of angiography in the state. Stage II models were developed in the same way for each stage I model coefficient. Thus, for example, the log odds of angiography for black versus nonblack persons in each state were also modeled as a function of region and angiography availability.

This approach recognized several sources of variation: within the same region, for a given level of angiography availability, states vary; within state, for a given set of patient characteristics, patient outcomes vary; and finally, variation remains after accounting for both patient and state characteristics. Differences in observed rates across states reflect all three sources of variation. The approach is similar to the empirical Bayes method, which recognizes two sources of variation, within units and across units; thus, empirical Bayes represents a special case of hierarchical modeling and the same types of shrunken estimates result. For example, the log odds of angiography in a particular state is a weighted combination of the intercept from the model that only includes patients from that state (stage I model) and the predicted value from the stage II model based on the region and availability of angiography in the state. The coefficient associated with the effect of race on angiography is a weighted combination of the coefficient from the stage I model and the

predicted value from the stage II model. As in empirical Bayes estimation, the degree of shrinkage is a function of the reliability of the within-unit estimate (here, within state) and the estimate of variation across states.

Interval estimates of parameter values from hierarchical models "quantify uncertainty," although they are not CIs.[11] Goldstein and Spiegelhalter (1996) illustrated this approach by reexamining the New York state CABG mortality data (see Chapter 1). They found very wide Bayesian intervals, which precluded definitive conclusions about most surgeons. For example, the analysis supported strong conclusions about whether rankings fell into the top or bottom half for only two of 17 surgeons. Green and Wintfield (1995, 1230) had criticized New York's CABG report because "in one year 46 percent of the surgeons had moved from one half of the ranked list to the other." Goldstein and Spiegelhalter (1996, 404) noted that "such variability in rankings appears to be an inevitable consequence of attempting to rank individuals with broadly similar performances." Furthermore (Goldstein and Spiegelhalter 1996),

> An over-interpretation of a set of rankings where there are large uncertainty intervals . . . can lead both to unfairness and to inefficiency and unwarranted conclusions about changes in ranks. In particular, apparent improvements for low ranking institutions may simply be a reflection of "regression to the mean."

Hierarchical models deal comprehensively and appropriately with the problem of multiple comparisons, as both point and interval estimates for each provider derive from all the data rather than just data for that particular provider. As Greenland (2000b, 920) notes:

> Giving the target parameters random components [as is done in hierarchical models] treats the problem [of multiple comparisons] with a global loss function quite different from that in classical adjustment: . . . modeling of the sort described here attempts to minimize estimation error by using additional background information, while classical methods only attempt to preserve global α-levels through purely arithmetic adjustment. It should come as no surprise, then, that critics of the latter find mixed [that is, hierarchical] modeling more acceptable.

In reanalyzing CABG mortality data from the Pennsylvania Health Care Cost Containment Council, Localio and colleagues (1997, 280) used simulations to demonstrate "the dramatic reduction in the number of false outliers with the use of hierarchical statistical models. The hierarchical models maintained adequate statistical power for detecting true departures from expected rates of mortality."

Hierarchical models rapidly become complex, requiring computer-intensive simulations to solve for parameter estimates, although computationally efficient approaches exist for conducting the simulations (Gelfand and

Smith 1990). New, easier-to-use software for personal computers is constantly evolving. The software package BUGS (Bayesian Inference Using Gibbs Sampling) is available free of charge from the United Kingdom's Medical Research Council Biostatistics Unit at the University of Cambridge Institute of Public Health (see http://www.mrc-bsu.cam.ac.uk/bugs/ for information on downloading and using the software). BUGS obtains solutions to the models using Markov Chain Monte Carlo simulation methods. This is a very powerful approach, although its use requires some statistical sophistication.

One advantage of simulation-based methods is that analysts can estimate more policy-relevant outcomes. Normand, Glickman, and Ryan (1996) did this in a study profiling hospitals for the HCFA Cooperative Cardiovascular Project in the early 1990s. Outcomes included the probability that hospital-specific mortality for average patients was at least 50 percent greater than median mortality, and the probability that the difference between risk-adjusted mortality (calculated for each hospital using a logistic regression model fit to the hospital's patients) and standardized mortality (predicted mortality based on a model developed from all patients) was large. Simulations enable relatively straightforward calculations of such statistics.

An Example Using Bayes Estimation

As an example of Bayesian methods, we simulated patient-level cost data and then used two approaches to estimate underlying parameters: shrunken estimates from a hierarchical model and averages calculated directly from the data. We illustrate two things: (1) Bayes estimates are more likely to "get it right" than traditional estimates and (2) how Bayes intervals and traditional CIs compare.

We assumed that the cost data were generated according to a slight enrichment of the simple hierarchical model described above. Specifically, each patient's cost was randomly sampled from a lognormal distribution with parameters that varied from provider to provider. For each provider, the parameters for the lognormal distribution were randomly sampled from a common normal distribution (truncated to be positive).[12]

We estimated parameters for the underlying distributions based on charge data for patients under age 65 admitted to Massachusetts hospitals in 1997 in DRGs 89/90 (simple pneumonia and pleurisy). Average costs per patient (more precisely, charges) were about $6,400, with an SD of roughly $4,000. We assumed hospitals' true mean costs varied from about 20 percent below to 20 percent above the average of $6,400.

For 25 hospitals, we generated data under two alternative assumptions about patient volume: first, assuming that each hospital treated 30 patients, often the minimum considered acceptable for producing profiles; and second, assuming that each hospital treated 100 patients, a relatively large number for a condition-specific profile. We generated five sets of simulated data under either scenario. From the simulated data, we estimated each hospital's mean two ways: (1) as the average of the data for the hospital and (2) using a

hierarchical model with noninformative priors (corresponding to vague prior knowledge) on unknown parameters (i.e., the mean and SD of the normal distribution and one of the parameters of the lognormal distribution). We ranked hospitals from high to low cost based on their mean costs estimated each of the two ways. We then compared the rankings to rankings based on their actual mean costs, which we knew because we knew the distributions from which the data were simulated. We report two measures: (1) on average, how far hospitals' ranks were from their true ranks (calculated as the average of the absolute value of the difference in ranks) and (2) how often hospitals' ranks were five or more ranks away from their true ranks (which would put them in a different quintile).

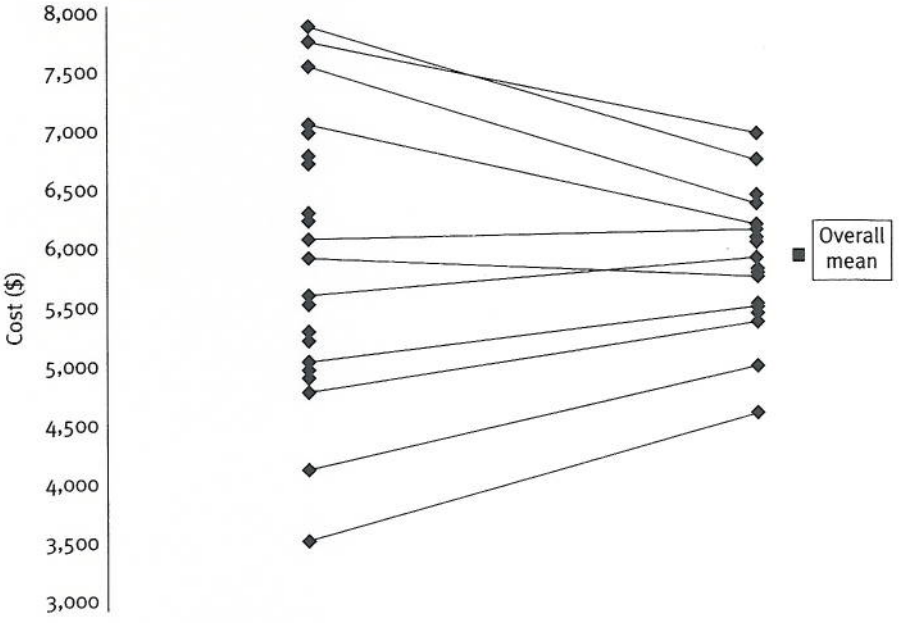
With sample sizes of 30, ranks based on the mean cost of a hospital's patients were, on average, 4.8 positions away from true ranks; more than 40 percent of the ranks were five or more positions away from true ranks. Ranks derived from the hierarchical model were, on average, 2.1 away from the true ranks; fewer than 8 percent of ranks were five or more ranks away from the true ranks.

We found similar results with sample sizes of 100, a size usually thought sufficient to "get things right." The ranks based on the mean cost of a hospital's 100 cases were, on average, 3.6 away from the true ranks; more than 35 percent of the ranks were five or more positions away from the true ranks. Ranks derived from the hierarchical model were, on average, 1.6 away from true; fewer than 5 percent of the ranks were five or more positions away from true ranks. Simulation results, however, probably overstate the real value of Bayesian estimation because we used the correct underlying model to make the Bayesian estimates. In real life, the hypothesized model is only an approximation of the underlying reality.

Figure 12.7 shows, for sample sizes of 30, the raw averages and shrunken estimates for the 25 hospitals from one replication of the simulation. (To avoid cluttering the graph, we connect only some of the pairs of estimates.) The shrinkage is evident. Estimates for hospitals at the extreme are "pulled" toward the mean, suggesting by how much raw averages overestimate differences among hospitals compared to the estimates from hierarchical models. Because all sample sizes are equal, differences in the amount of shrinkage are caused by differences in the distribution of hospitals' cost data, especially the influence of outliers. Consider the two most expensive hospitals. The most expensive hospital's shrunken estimate was below that of the second most expensive hospital. The explanation is that two very expensive outliers caused the most expensive hospital's high average costs. In contrast, the second most expensive hospital had many cases with relatively high costs but no extreme outliers. The Bayesian estimates discount outliers and shrink the estimates more when outliers drive the raw averages.

Figure 12.7 also demonstrates that hierarchical models do not necessarily shrink all estimates toward the overall mean. In fact, the hierarchical model pulls the tenth most expensive hospital's cost away from the mean.

FIGURE 12.7
Average Cost (left) and Bayes-Estimated Mean Cost (right) by Provider*



* Calculated from simulated data for 30 patients at 25 hospitals.

Examining the distribution of this hospital's costs is informative. This hospital had many inexpensive cases, but one very costly outlier pulled up the average. After reducing the effect of this outlier, average cost fell, and the shrunken estimate pulls this "average with outlier effect reduced" toward the mean.

For five hospitals, Figure 12.8 shows the estimated means as well as CIs and Bayesian probability intervals. Figure 12.8 makes two points. First, Bayes estimates pull extreme averages toward the overall sample mean, which is slightly under $6,000; second, Bayesian intervals are frequently wider than the CIs.

Hierarchical modeling provides an attractive framework for estimation when profiling providers. Shrunken estimates appropriately adjust for the influence of outliers and the increased unreliability associated with estimates from smaller samples. Furthermore, the Bayesian probability intervals better reflect the uncertainty associated with estimates than do traditional CIs. Nevertheless, hierarchical models have generally not been used to profile provider performance outside of research settings. One criticism is the extent to which results are based on the underlying probability models, although Greenland (2000c, 164) noted:

> Every inferential statistic (such as a p-value or confidence interval) is model based, in that some set of constraints (i.e., a model) on the data-generating process must be assumed in order to derive tests and estimates of quantities of interest. . . . Multilevel modeling is

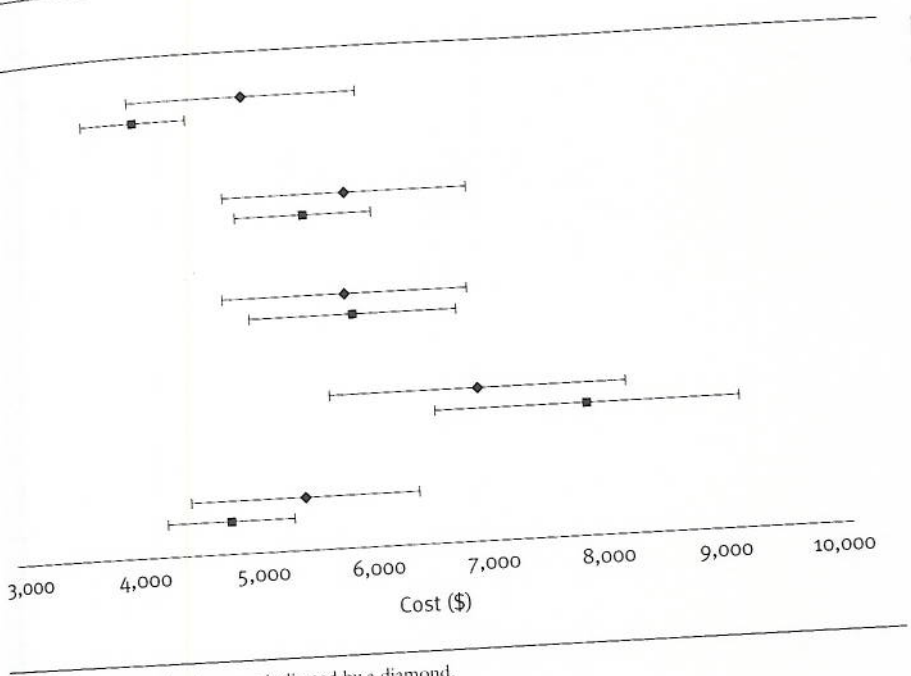


FIGURE 12.8 95 Percent Bayesian Probability Interval* and 95 Percent Confidence Interval** for Simulated Data For Five Providers

* Top interval of each pair, mean indicated by a diamond.
** Bottom interval of each pair, mean indicated by a square.

distinguished only by its unfamiliarity, which obliges one to make more effort to explain the model. But multilevel modeling need not involve stronger assumptions than ordinary modeling, and in fact provides an opportunity to use weaker assumptions.

A more serious problem is that hierarchical models "require substantial statistical sophistication to implement properly" (Shahian et al. 2001, 2162). Hierarchical models nonetheless offer substantial advantages, and ever-easier methods for implementing these models will likely continue to appear.

Comparing Outcomes over Time

Chapter 12 thus far has concentrated on cross-sectional comparisons—information relating to a single time period. However, a powerful profiling tool involves examining changes over time, so-called "longitudinal" analyses. As Donald M. Berwick, M.D. (1996, 4), a leading health care quality-improvement expert, observed, "Pick a measurement you care about, and begin to plot it regularly over time. Much good will follow."

Plotting measures over time highlights change. For simplicity, suppose that all patients have the same level of risk. Suppose that we examine the problem rate in two hospitals in year 1. Hospital A has a problem rate of 3 percent, with a 95 percent CI from 2 percent to 4 percent, whereas hospital B's rate is 5 percent, with a 95 percent CI from 4 percent to 6 percent. From the classical "hypothesis testing" perspective, we can reject the hypothesis that the underlying problem rates at the two hospitals (P_A and P_B) were identical in year

1 in favor of the alternative that hospital B's rate was higher. This conclusion does not mean that hospital B's problem rate will be higher than A's next year. Furthermore, even if in year 2 hospital A's problem rate is statistically significantly higher than B's, that does not mean that the assessment of which facility did better in year 1 was wrong. Hospital B could have improved.

However, provider profiles are useful chiefly to the extent that they reflect a persistent reality. As noted earlier, Green and Wintfeld (1995, 1230) criticized New York state's CABG mortality report. They said:

> The usefulness of the risk-adjusted data was also limited in that surgeons' rankings during two years of the study offered few clues about their position in the subsequent year ($R^2 = 4.9$ percent). . . . The fact that surgeons' performance ratings fluctuate so much from year to year means that by the time the data are published, users of the report can have little confidence that the ratings are still applicable.

Green and Wintfeld thus speculate that real differences in comparative performance are outdated by the time they become publicly available. However, Goldstein and Spiegelhalter (1996) provide a more fundamental critique, suggesting that such large changes in rank are likely when true differences in provider performance, even if real and stable over time, are small compared to random "noise."

For the NSQIP (see Chapter 8), Khuri and colleagues (1998) illustrated a way to portray *O/E* ratios over time. Such presentations suggest that providers' performance varies from year to year. From the numbers alone, however, we cannot know how much variability results from some providers improving more than others (i.e., last year's data are outdated) and how much from randomness (i.e., noise overwhelms the "signal"). In-depth study is required to disentangle these different possibilities. Nevertheless, longitudinal plots can reveal when providers' ranks change dramatically from one year to the next. When big yearly changes are common, both public reporting and decision making should be restrained. In particular, report cards should not list providers in rank order of their measured performance, as this reinforces the impression that the figures are reliable. Managers should think twice before disrupting provider-patient and network-provider relationships over findings that may be transitory, even if real.

Given the various methodological concerns and resulting questions about interpretation, profiles should be employed only where they are likely to be useful. For example, if last year's findings differ from this year's because relative quality can change rapidly among providers, profiling data will be most valuable for quality improvement and less useful to large purchasers and individual consumers of health care services aiming to select providers for the future. Even when longitudinal analyses show stability over time, legitimate concerns remain over whether consistently poor performers are the victims of inadequate risk adjustment. Future research must identify profiling infor-

mation that is relatively stable over time and distinguish it from figures that fluctuate without obvious explanation. With most current profiling initiatives, random noise is a major consideration, as are unmeasured differences in patient risk. Small sample sizes for individual providers also raise concerns. These factors limit the inferences that can be drawn from practice profiles. Longitudinal plots often provide a sobering reality check to the evaluation process.

Despite methodological concerns, profiles are increasingly generated and used as important tools in ensuring health care "value"—a melding of cost and quality. Comparing patient outcomes across providers can be valuable, but much depends on how the profiles are used. Given the state of the art, however, relying on such profiles alone to make all-or-nothing business decisions (e.g., withdrawing business from outlier providers) is inappropriate. In this context, profiles are likely to generate (often well-founded) criticism and heighten adversarial relationships among providers, payers, and policymakers. Similarly, if such profiles are disseminated to a public unaware of the need for cautious interpretation, further controversy may erupt, impeding opportunities for useful dialog and improvements. If providers are given profiles without education about how to use them productively to identify areas for improvement, the information will likely be ignored.

Profiles comparing patient outcomes are most valuable in an environment of cooperation and collaboration, with incentives for learning and improvement. With increasing competitive pressures, however, this ideal environment may be more pipe dream than tangible reality.

Notes

1. The coefficient of variation, or CV, defined for a nonnegative variable as s/μ, is a useful summary measure. Hospital costs for a specified type of hospitalization often have a CV of around 1. In looking at total costs next year for a heterogeneous population (with many zeros and a few extreme outliers), the CV may be 4 or larger.
2. Recognizing that averages based on moderate sample sizes are only approximately normal, it makes sense to avoid the "appearance of precision" implied by using intervals with half-widths equal to 1.96 * SE. Thus, we use 2 * SE.
3. s is calculated as the weighted average of the SDs at each hospital, as described above. If the data were distributed normally, $\overline{Y}_A$ would be in this interval about 95 percent of the time. With highly skewed cost data and modest sample sizes, the probability is lower. As a result, under the hypothesis of no differences among providers, having more than 5 percent of "normative" providers fall outside these bounds is not surprising.
4. The calculations of Table 12.3 are only approximate. Especially when P is near zero, a reasonable 95 percent CI would not be centered

at $\hat{P}$. For example, after observing 5 deaths in 500 patients (1 percent), the modified Wald method that Agresti and Coull (1998) recommend yields the 95 percent CI 0.4 percent to 2.4 percent. Note, however, that the half-width of this interval is 1 percent, as suggested by the calculations in the table. For a clear discussion of these issues, see " 'Exact' Confidence Intervals Are Not Exactly Correct" at http://www.graphpad.com/articles/CIofProportion.htm.

5. A better CI—of approximately the same width—would be 5.4 percent to 17.6 percent.
6. Odds—the ratio of the probability of an event to the probability that it will not happen—are much used in the world of betting. For example, an event has 3-to-1 odds (i.e., odds = 3) when it has a $P = 0.75$ likelihood of happening and $P = 0.25$ chance of not occurring.
7. This expected rate is identical to that from a model that predicts probability of death from the whole population using the single predictor of high versus low risk.
8. Greenland (2000c) gives an excellent nontechnical description of the principles of multilevel modeling. McNeil, Pedersen, and Gatsonis (1992) also provide a nontechnical description of hierarchical models in the context of provider profiling. Shahian and colleagues (2001) discuss problems with traditional approaches and the advantages of hierarchical models as part of a review of cardiac surgery report cards. Normand, Glickman, and Gatsonis (1997) provide a technical discussion of statistical methods for profiling providers.
9. For ease of exposition, we ignore here the interesting observation that "you can load a die, but you can't bias a coin" (Gelman and Nolan 2002, 308).
10. Traditional thinking held that statistical estimators should be unbiased (i.e., the difference between the expectation of the estimator and the parameter being estimated equals zero). However, overall error, measured by the mean square error (MSE) has two components: bias in the estimated parameter and the spread of individual data points around the estimated parameter. Stein estimators minimize MSE but are biased. For profiling, it is probably fine to accept a small amount of bias if that leads to a smaller MSE.
11. A 95 percent CI is an interval that has a 95 percent chance of covering the parameter of interest; if data were resampled from the population 100 times and 95 percent CIs constructed, about 95 of these 100 intervals would include the underlying parameter. Note that in this interpretation, the intervals have a chance of covering a fixed parameter. Ninety-five percent CIs are often interpreted incorrectly as being the interval in which there is a 95 percent chance the parameter value falls. In this incorrect interpretation, the parameter is viewed as having a chance of falling into a fixed interval. The distinction is between the

chance the interval covers the parameter (which is what a CI is) and the chance that the parameter falls in the interval (the way in which a CI is incorrectly interpreted). A non-Bayesian framework assumes that the parameter value is fixed. Hence, considering the chance that the parameter value falls in an interval makes little sense. A Bayesian framework creates a probability distribution for parameters; therefore, considering the chance the parameter lies in some interval makes sense. This is the interval determined in a Bayesian analysis.

12. A lognormal distribution with parameters μ_j and σ^2 has mean $\exp(\mu_j + \sigma^2/2)$ and variance $\exp(2\mu_j + 2\sigma^2) - \exp(2\mu_j + \sigma^2)$. Here, we supposed σ^2 was fixed and that for each provider the parameter μ_j was generated according to a common normal distribution. Then, we generated patients' costs according to a lognormal distribution with parameters μ_j and σ^2 and took the true mean for a provider to be $\exp(\mu_j + \sigma^2/2)$.