Improving the Statistical Approach to Health Care Provider Profiling

Cindy L. Christiansen, PhD, and Carl N. Morris, PhD

This paper reviews and compares existing statistical methods for profiling health care providers. It recommends improvements that are based on the use of better statistical models and the adoption of more realistic, medically based criteria for judging the performance of health care providers. Unlike most profiling methods, the proposed hierarchical models allow the probability of acceptable provider performance to be calculated; thus, they can answer such questions as, "What is the probability that a given hospital’s true mortality rate for cardiac surgery patients exceeded 3.33% last year?" The commonly encountered problems of regression-to-the-mean bias and small caseloads can be handled by using hierarchical models to extract more information from profiling data.


From Harvard Medical School and Harvard Pilgrim Health Care, Boston, Massachusetts; and Harvard University, Cambridge, Massachusetts. For current author addresses, see end of text.

For reports on the performance of health care providers to be effective, profiling must be done using the best statistical methods. Commonly used profiling methods often contain some of the following deficiencies. They ignore important relevant information. They use statistical standards where medical standards would serve better. They use the probability that the observed outcome is extreme, assuming that a hospital's true performance is acceptable to identify providers with extreme rates; this is not the probability that the medical unit's true mortality rate exceeds a given standard. (The true mortality rate is the rate that would have occurred if the hospital had served a very large number of patients.)

The 1993 report on coronary artery bypass graft surgery from the New York State Department of Health (1) listed mortality data (deaths before leaving the hospital) and profile statistics for 31 hospitals. These profiles aimed to identify hospitals that had excessively high or low mortality rates associated with coronary artery bypass graft surgery. In this article, we review 3 of the 31 hospitals to see how profile results improve with the use of more information. Two of the hospitals were chosen for their high observed mortality rates (1 had a mortality rate that was substantially higher than the 1992 statewide rate of 2.78%), and 1 was chosen for its low mortality rate. (It is important to keep in mind that the performance of the hospitals may have changed since 1992.)

Hierarchical models use information from the available data obtained from all health care providers being examined. These models are so named because they apply to situations with two or more levels of random variation. In the mortality rate example to follow, the first level of the hierarchy specifies a distribution for the random number of observed deaths at a given hospital and the second level specifies possible distributions for the true mortality rates in several hospitals. To use terminology from analysis of variance, level 1 in the hierarchical model concerns the variation of rates within providers and level 2 concerns the variation between true rates of the hospitals.

We urge the use of medical standards that specify the largest or smallest medically acceptable true mortality rate in the setting being profiled. How standards are set depends on their purpose; standards meant to encourage quality improvement, for example, may differ from standards meant to distribute pay incentives. Major improvements to profiles will result from the use of medically appropriate performance standards.

Case-mix adjustments are made in almost all profile analyses to account for the differences in provider performances attributable solely to differences in the populations served. Hierarchical models accommodate these crucial adjustments. In addition, standard profiling procedures typically ignore units with small caseloads, such as those with fewer than 50 patients. This practice gives little information on the performance of low-volume providers and can provide unfair gaming opportunities (for example, a hospital with a very high mortality rate for 49 patients might refuse to admit further patients in the given category). Hierarchical models require no minimum sample size for a particular health care provider, provided that the ensemble of all providers analyzed has adequate data. Ensemble data are used to correct for regression-to-the-mean bias.

In this paper, we review standard statistical profiling methods, showing how successively better results are obtained as more information is included. We recommend the use of hierarchical models to extract ensemble information and advocate the use of more directly relevant standards.

The advantages of a hierarchical approach are as follows: 1) The probabilities of performance standards are calculated, 2) comparisons of units are
based on medically relevant standards, 3) regression-to-the-mean bias is removed, and 4) providers with small sample sizes remain in the analysis. The benefits of hierarchical modeling apply not only to the mortality data considered here but also to profiles of patient satisfaction, referral rates, and other outcomes.

Profiling Data

We estimated the true mortality rates associated with coronary artery bypass graft surgery for 31 hospitals in New York State (1) by using data on the number of patients, the number of deaths, and the case-mix difficulty of the patients treated at each hospital. For this analysis, we focused on data for two hospitals (H1 and H2) that had high observed mortality rates.

The simplest profile analysis would compare the number of deaths in each hospital. Because 22 deaths occurred at H2, it seems to be a much worse hospital than H1, at which 3 deaths occurred. However, the number of patients served should also be considered: that is, mortality rates, not raw counts, should be analyzed. The disparity in the number of deaths can be completely explained by the caseload of 484 coronary artery bypass graft procedures at H2 compared with 67 at H1. The mortality rates of 4.48% (3 of 67 patients) at H1 and 4.55% (22 of 484 patients) at H2 are almost indistinguishable. This improvement makes a fairer comparison by adjusting for caseload while retaining the basic concept of comparing the number of deaths. The improvements stem from obtaining and using appropriate additional information.

An even better approach accounts for case-mix differences. Data from the New York state study (1) show that the patients who had coronary artery bypass graft surgery at H2 were less healthy than those at H1. On average, patients at H1 had 51.1% of the risk for death of all patients who had coronary artery bypass graft surgery (expected mortality rate, 1.42% for the case mix of patients at H1 compared with 2.78% statewide; 1.42 ÷ 2.78 = 0.511). When we adjusted for this, the 3 deaths at H1 equivalently resulted from 34.2 (0.511 × 67) procedures done on patients with an average case mix. The risk-adjusted mortality rate is therefore 8.77% (3 ÷ 34.2) at H1. The risk-adjusted mortality rate at H2 was 5.77%. (With many deaths and an exceptionally healthy patient case mix, it is possible for the relative risk case-mix adjustment method to produce adjusted rates exceeding 1.00.)

Risk adjustments contribute vitally to reducing unfair profile evaluations. The need for risk adjustment has led to vigorous research on ways to ac-

count for case-mix differences (2–8). For our purposes, we used the case-mix data and the risk adjustment methods as they were used in the report from New York State (1).

Models and Tests of Statistical Significance

Developing good profiling procedures requires specifying probability distributions for the observed outcomes and choosing the standards of acceptable care that define the hypotheses to be tested. Standards that are based on input from medical professionals and from users of the profiles will be the most useful and the most meaningful. When statistical convention alone determines these choices, they are likely to produce inaccurate conclusions and lead to poor decisions.

Probability Distributions

Once the hospital rates have been adjusted for case mix, a probability distribution is needed to perform a statistical test. The observed count is governed by the hospital's true mortality rate; here, it is the rate that could be observed only if the hospital had treated a very large number of patients with the average case mix. The true mortality rates are unknown and must be estimated from the data.

The Poisson distribution is appropriate for these data because the probability of death after coronary artery bypass graft surgery was small (2.78% statewide). An alternative, the binomial distribution, is well approximated by the Poisson distribution in this case. If the expected number of deaths is very small or if some individual probabilities are large, then normal, Poisson, and binomial assumptions may not be valid and an alternate calculation may be necessary (9).

Commonly Used Decision Criteria

Profile analyses often require tests of the null hypothesis that a provider’s true mortality rate equals the average rate for all providers. The hypothesis is tested at a specified significance level. Following this convention, the New York State report (1) used the statewide mortality rate of 2.78% as the standard and set the significance level at 0.025. This hypothesis is not very useful: Taken literally, it means that if the true hospital mortality rates differ even by tiny amounts (which one would expect), many of the hospitals would have true rates that exceed the population mean.

P Values

When a distribution and a standard are specified, the P value can be calculated (10, 11). A normal approximation to the Poisson distribution will work
poorly for the profiles of New York State hospitals in which coronary artery bypass graft surgery was performed, because fewer than 10 deaths were expected at many of the hospitals. The $P$ value for H1 is the probability of observing 3 or more deaths (because 3 deaths occurred at H1), assuming that H1 performed with a true mortality rate of 2.78% for patients with an average case mix. Had H1 performed at this average rate, 0.95 deaths would have been expected. A normal approximation produces a $P$ value of 0.018. Small $P$ values, such as this, identify high true mortality rates; H1 would therefore be identified by this approximate calculation. However, the exact $P$ value based on the Poisson distribution is 0.072 and is too large to identify H1 as a poor performer. Profile procedures that use normal approximations result in incorrect profile estimates if the approximation is inaccurate. Errors such as this are unnecessary; many statistical computer packages make it easy to calculate exact $P$ values for the Poisson and other common distributions.

Hierarchical Bayesian Models for Profile Analyses

A $P$ value computed to test the hypothesis that a hospital has performed adequately differs from the probability that the hospital has performed adequately. The $P$ value for H1 is 0.072, but this is not the probability that H1 performed acceptably. Hierarchical models and Bayesian reasoning allow the probability of acceptable performance to be calculated.

The probability of acceptable performance cannot be calculated unless a prior probability—that is, the probability distribution that describes the variation of a hospital’s true mortality rate before the data for that hospital were observed—is first specified. If expert opinion is unavailable to specify this prior distribution (or if this opinion is not trusted), the distribution is usually chosen to reflect extreme uncertainty about the true rate. In these cases, the probability of adequate performance will often be close to the exact $P$ value (12). One can estimate this distribution by using a hierarchical model (13) and the data for all hospitals. By using the data from all the hospitals, we include more information in the analysis and can expect better estimates.

Much recent research has centered on hierarchical statistical methods for normal, Poisson, binomial, and other distributions that could help in medical profiling (14–16). In a hierarchical model, the first-level distribution for the observed data and the second-level prior distribution (also called the population distribution) for the true mortality rates are used to compute the probability that a hospital has performed acceptably. We recommend that $P$ values be replaced by this probability.

The hierarchical model that we fitted to the Poisson data (by using case-mix information and Poisson Regression Interactive Multilevel Modeling software [17]) has been discussed elsewhere (18). The fitted prior distribution has a mean mortality rate ($\pm SD$) of $2.78\% \pm 0.61\%$. Almost all of the hospitals had true mortality rates between 1.56% and 4.00%, within 2 SDs of the mean.

We then used each hospital’s observed mortality rate to construct an estimate of a hospital’s true mortality rate. Figure 1 shows the observed mortality rates for the 31 hospitals and the estimated true mortality rates. The estimates of the 31 mortality rates done using a hierarchical model are less variable than the observed rates. Estimates of the true
mortality rates are based on shrinkage coefficients, which are values between 0 and 1. The shrinkages determine how far the observed rate at the hospital moves (that is, shrinks) toward the population rate of 2.78% (Figure 1) to produce the best estimate for that hospital. Shrinkages close to 1 correspond to small caseloads, like that of H1, for which the best shrunken estimate is close to the overall mortality rate of 2.78%. Larger caseloads, such as that of H3 (which also had a low mortality rate), have smaller shrinkages; therefore, the estimate is relatively close to the hospital’s observed rate. The average shrinkage for the 31 hospitals is 65%, indicating that the true (unobserved) hospital rates vary less around the mean than the observed rate estimates vary around their true hospital means.

Each shrinkage coefficient estimates how much regression-to-the-mean is expected at a hospital. The mortality rate for future patients at a hospital would be expected to regress away from the observed rate and toward the population rate in proportion to the estimated shrinkage. Mortality rate estimates that are unadjusted for regression to the mean are biased in the sense that the observed mortality rate estimate will usually be farther than the true rate from the population mean. Hierarchical models provide more accurate estimates by making the appropriate adjustment. Our example treats all hospitals as having the same mean because the rates were standardized to have this population case mix, but hierarchical models could be fitted by shrinking each observed estimate toward a different fitted value.

**Better Criteria and New Standards**

Profiling criteria often compare each health care provider with the average performance of all providers. More appropriate performance standards can be developed that are based on acceptable medical practice and on costs and benefits to the public. These standards must be set jointly by medically informed policymakers, the users of the profile, and others parties. An accreditation agency, for example, might evaluate a provider’s performance by different standards than would patients. Once the statistical analysis has determined the probability distribution for each provider’s true rate, little additional effort is required to compute the probabilities for various criteria.

If the maximum acceptable mortality rate for coronary artery bypass graft surgery were set at 3.33% for an average case mix (about 20% greater than the mean of 2.78%), hospitals whose true mortality rates are greater than the average are acceptable, provided that these rates do not exceed 3.33%.

With this less stringent standard, a hospital’s performance might now be judged as substandard if the probability that it has exceeded the standard is greater than 0.50 (that is, it is more likely than not that the true rate is >3.33%). These two values, the standard for acceptable care and the minimum probability for compliance, should ideally be determined with the collaboration of all knowledgeable and interested parties.

**Calculating the Probability of Poor Performance**

Figure 2 shows the combined use of hierarchical statistical models with the suggested performance criteria. The population rate, the chosen standard, and the distributions that specify uncertainty about the true rates at H1, H2, and H3 are indicated. The mortality rate at H2 is very likely to be between 2.7% and 4.9% (and therefore has a high probability of exceeding the earlier standard of 2.78%). The true mortality rate at H2 has a probability of 0.75 of having exceeded 3.33% because 75% of the area under the curve of H2 lies to the right of 0.033. Thus, H2 probably had an excessively high mortality rate. The probability that the true mortality rate at H1 exceeded 3.33% is 0.28. Even though H1 had a higher observed rate, it was more likely to have performed acceptably for the given standard. (Of course, even a 28% chance of receiving poor care may not be satisfactory in some situations.) The probability that H3 performed poorly is negligible.

The second highest probability of exceeding 3.33% among the 31 hospitals was found in a fourth hospital with a probability of 0.48. Although that hospital barely meets the specified standard, the 48% probability of exceeding 3.33% provides directly interpretable and relevant information to the profile user. P values are not directly interpretable this way.
Limitations of Hierarchical Models

Statistical methods, including methods for hierarchical models, cannot provide better information than the available data warrant, nor can they overcome the problems of inadequate case-mix measures. Although hierarchical models provide more accurate estimates than standard methods, they do so only if they are well specified. Poorly specified statistical models produce misleading results. Statisticians also must ensure that the chosen model, whether simple or hierarchical, fits the data (19).

Although hierarchical models are only slightly more advanced than ordinary regression models, they are relatively unknown to most profile users and are not always well understood by them. Interpreting probability models requires careful thought and knowledge of their underlying assumptions and limitations (20). Until hierarchical models are commonly found in medical and health service curricula, we recommend that they be used only with the help of expert statisticians.

Conclusions

Hierarchical models are powerful statistical tools that have been the focus of much development in recent years. They have been used in medicine, education (21–23), and wherever many estimates must be made. In addition to the standards that we discussed, provider rankings (21) and other standards are often used. Commercial and publicly available software is available to fit these models (17, 24, 25).

Hierarchical models are expected to increasingly appear in medical research and journal articles (26). These models are particularly appropriate for profiling applications because several providers are usually profiled simultaneously. The use of ensemble information in hierarchical models allows the probability of each provider's performance to be calculated, yielding directly interpretable results and relevant answers to the questions being asked about health care quality, access, and satisfaction. Careful use of such models, together with improved standards, will make medical profiles more accurate and more meaningful for health care policy and decision making.

Grant Support: In part by Agency for Health Care Policy and Research grant RO1 HS 07118-01 and the Department of Veterans Affairs, Management Science Group.

Requests for Reprints: Cindy L. Christiansen, PhD, Harvard Medical School and Harvard Pilgrim Health Care, 126 Brookline Avenue, Suite 200, Boston, MA 02215.

References
