We have entered a new era of medicine, in which physicians are no longer granted an assumption of excellence. Studies have demonstrated marked variability in the practice of medicine and revealed abundant opportunities for improvement in the care of patients. These studies have accentuated the pressure exerted by payers and the public on the profession to assume greater responsibility for delivering high-quality care.

In this era of accountability, ratings and rankings proliferate as the public and payers seek information about the performance of physicians, hospitals, and health plans. The consequences of these rankings can be profound as marketing departments and the popular press seize the results. Hospitals ranked highly tend to tout their status in advertisements, whereas those with less favorable ratings hope to avoid being identified as hazards.

The interest in rankings is particularly strong in cardiovascular medicine. Cardiovascular diagnoses represent a substantial proportion of high-volume and high-cost admissions to hospitals. Statewide and national efforts have focused on institution- and physician-specific outcomes after cardiovascular procedures and acute myocardial infarction. In several areas, such as bypass surgery, mortality rates are the most common basis for quality-of-care rankings. In New York, bypass surgery ratings of individual surgeons based on mortality rates are highly publicized in the press. Studies suggest that these public ratings have stimulated improvements in the care and outcomes of patients, although the evidence is indirect.

Because patients are not randomly allocated to different physicians or hospitals, the comparison of outcomes among various sites is challenging. Differences in referral patterns and catchment areas may result in different populations at each institution. Consequently, the comparison of crude outcome rates may be misleading. The same mortality rate may be interpreted differently depending on the profile of the patients. An operative mortality rate of 2% for bypass surgery may be better than expected at a hospital that operates on older, unstable patients and worse than expected at a hospital that accepts only younger, elective patients. Accounting for differences in patient characteristics is the key to an appropriate comparison.

In the background of these attempts to compare performance among institutions or individuals stand mathematical regression models that are intended to “level the playing field.” In particular, the logistic regression model has emerged as an accepted method of analysis. This type of model is appropriate for the evaluation of a binary outcome such as mortality, can provide an estimated risk for each patient, and is easily calculated with contemporary software. Once the logistic regression model is developed, it is possible to characterize the clinical status of a group of patients by calculating the mean of their individual predicted rates of a particular outcome. This calculation gives an overall prediction or expected outcome rate based on the patients’ characteristics. Thus, for a given group of patients, it is possible to compare the actual outcome rates with the predicted outcome rates.

The validity of this approach depends on the mathematical model. However, the adequacy of the model is not always readily apparent. Data are entered and results are produced, but the computer outputs may not reveal problems with the models unless specific diagnostic calculations are performed. Unfortunately, most physicians lack the training to critique the model. Most, but not all, physicians seem to accept this approach. A survey of cardiologists in New York revealed that 67% found the comparison of cardiac surgeons by the adjusted mortality rates to be “very accurate” or “somewhat accurate.” Conversely, a substantial minority of cardiologists expressed doubts about the methodology, with about a third reporting that the findings are “not accurate at all.”

Given the importance of accountability and the frequent use of these models to rate physicians and hospitals, we need to focus attention on them and evaluate the impact of various approaches. Many issues remain unresolved. What are the minimum standards that make a model acceptable to compare outcomes? How often should a model be updated or revised? How confident can we be that the models separate the real differences from random variation? What sample sizes are required? What are the most efficient models that are least susceptible to intentional manipulation, or “gaming?” These models are too important to remain black boxes to the profession. We need to increase literacy regarding risk-adjustment approaches and to pursue methodologically rig-
orous research to address the gaps in our knowledge about them.

In this issue of Circulation, Ivanov and colleagues address the important issue of model selection and revision by focusing on coronary artery bypass graft surgery. This issue is very relevant for institutions and organizations that are using a model developed several years ago to produce an estimate of risk-adjusted outcome. Does the model need to be updated? What is gained by changing the model?

To address this issue, the authors consider 3 options. In the first approach, they use a model that was developed in a similar population without any change ("ready-made" model). In the second approach, they use the variables from the previous model but recalculate the equation with their own data to generate new coefficients. The new coefficients are presumably more appropriate to their own population (as opposed to the population in which the model was originally developed) and will improve the predictive ability of the model without the need to change the approach to data collection. Finally, in the third approach, they develop a new model with the possibility of using new variables.

The authors use standard approaches to determine the best logistic regression model. First, they compare the calibration of the models, a measure of how well the predicted model fits the actual data. A good model should predict mortality at all levels of severity. Regression models may not accurately predict the risk of death for patients who are at either end of the spectrum of disease severity. Models have a tendency to predict a higher risk of death than is observed for the patients with the least severe illness and a lower risk of death than is observed for the most ill patients. A model that does not fit the actual data at all levels of severity is unsatisfactory and needs modification.

The use of a model that does not fit actual data can favor hospitals or physicians that admit certain types of patients. Consider 2 hospitals that have identical performance. The hospitals, by virtue of their sizes and referral areas, admit very different types of patients. One hospital admits mostly patients who are not very ill, whereas the other hospital admits patients who have catastrophic illness. A poorly calibrated model can create the perception of a very different performance by the 2 hospitals. The hospital that cares for mostly very-low-risk patients may have a predicted mortality rate higher than the observed mortality rate because the model tends to overestimate risk at the low end of severity. Meanwhile, the hospital that cares for the most ill patients may have a predicted mortality rate that underestimates the risk of mortality, leading to a higher observed than predicted mortality rate for this hospital.

Because the calibration of the model can have such a marked effect on performance measurement, it should be evaluated and reported for each model. Graphical displays of the model fit are useful but not quantitative. Goodness-of-fit tests are intended to determine whether the observed data deviate significantly from the fitted model. Unfortunately, goodness-of-fit tests often have relatively low power to detect model problems. If this test demonstrates poor fit, then the model is clearly not calibrated. However, the test is able to detect only marked conflicts between the fitted model and the data, and problems that are not gross may be overlooked.

In the Ivanov study, the investigators found that the existing model derived on a sample of patients from 1991 to 1993 did not have good calibration, with a significantly poor fit between predicted and observed results. The newer models had acceptable calibration. These results illuminate one of the issues in the evaluation of models. In this case, the poor calibration may not indicate an inferior model. Rather than indicating a model that fails at the ends of the spectrum of the patients’ clinical severity, the poor calibration in this case may be principally the result of improvement over time in surgical outcomes. All hospitals may have decreased their mortality rates. If so, a good model from 5 years ago would tend to predict higher mortality rates than are currently observed. The models based on older data would overestimate the contemporary mortality risk. This improved performance in outcomes would degrade the overall calibration of the model using current data but would not necessarily affect its utility in assessing the relative performance of the hospitals if the lack of calibration occurs equally across the spectrum of patient severity of illness.

Similarly, a national model that uses data from a single hospital or group of hospitals may not be well calibrated if their overall performance is different from the sample for which the model was derived. The lack of calibration may indicate that the single hospital performs better than expected. This information is important, and it would be unfortunate to lose this perspective by recalibrating the model.

All models should be calibrated when they are developed. However, the value of calibration when it is applied to other samples depends on the purpose of a model. For a model that is intended to give a good estimate of an outcome, such as for patient counseling, calibration is critical. If the model is intended to provide information about the relative performance of institutions or provide a comparison over time, the lack of calibration of a ready-made model in another data set may indicate an important difference in performance.

The investigators also evaluated the discriminative ability of the various models. Discrimination measures the model’s ability to separate patients with different outcomes. In the Ivanov study, the investigators determined how well the model differentiated those who survived from those who died. In a model with perfect discriminant ability, all the patients who died would have a higher predicted mortality than the patients who survived. They found small but statistically significant improvements in the discriminant ability of the recalibrated and new models compared with the ready-made model. The impact of such small differences is not known.

Fortunately, the authors also looked at the practical impact of the differences in calibration and discrimination by assessing the change in surgeon rank with the newer models. They found surprisingly little change for the effort. Among the 14 surgeons evaluated, no change in rank occurred among the top 7 in any of the 3 models. With recalibration, only surgeons ranked 8 and 9 by the ready-made model switched places. With the new model, the surgeon ranked 13 with the
ready-made index moved up 2 places, with numbers 10 and 11 moving down a place.

Despite this relatively unimpressive change in the rankings, the authors conclude that an existing model should be episodically recalibrated and compared with newer models. The recommendation is vague, perhaps indicating uncertainty about the best course of action. It seems prudent to examine these models periodically, but several questions remain. When should an old model be revised or replaced? What are the incremental benefits and costs? How should the model’s principal purpose affect its assessment? If the sole purpose of the model is to counsel the patient, then calibration and discrimination are paramount. The model should be updated frequently to provide the patient with the most current estimate given the evolution of medical and surgical techniques. If the model is meant for comparison, then updating the model will lose the historical perspective.

Although this article makes a valuable contribution in directing attention to the value of recalibrating or remodeling, there are also many other important issues that deserve our attention with regard to risk-adjustment models. Even the best models explain only a relatively small proportion of the variability in outcomes. It remains unknown whether these risk-adjusted mortality rates are an indicator of quality. Different models can result in profoundly different impressions of quality among institutions.10,11 Which models, if any, are best at identifying quality of care? Even if the model were appropriate, what sample sizes would be required to distinguish the signal from the noise? Another important issue regards the reliability of data collection and the possibility of gaming the model by increasing the severity through more complete documentation. The cost of data collection is also becoming increasingly important, leading to a greater interest in parsimonious models, or ones that can be derived solely from existing data information systems.

The shift toward accountability should help patients, but the methods of assessment must be scientifically sound and the choices about them defensible. Risk models will continue to have utility, but the temptation to use available but unproven approaches to assess quality must be resisted. The study by Ivanov et al has directed attention toward this practical issue of model selection that confronts many organizations. More research is needed to determine the validity of risk-adjusted outcomes as a measure of quality and the best approaches to account for the nonrandom allocation of patients across sites. For now, given the state of the art, it seems most prudent to use the comparative outcome information from risk-adjustment models primarily to stimulate benchmarking efforts and generate hypotheses about how to improve care.

References

Key Words: Editorials risk factors bypass