Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation

Harlan M. Krumholz, Ralph G. Brindis, John E. Brush, David J. Cohen, Andrew J. Epstein, Karen Furie, George Howard, Eric D. Peterson, Saif S. Rathore, Sidney C. Smith, Jr, John A. Spertus, Yun Wang and Sharon-Lise T. Normand

The online version of this article, along with updated information and services, is located on the World Wide Web at:

http://circ.ahajournals.org/cgi/content/full/113/3/456

Subscriptions: Information about subscribing to Circulation is online at
http://circ.ahajournals.org/subsriptions/

Permissions: Permissions & Rights Desk, Lippincott Williams & Wilkins, 351 West Camden
Street, Baltimore, MD 21202-2436. Phone 410-5280-4050. Fax: 410-528-8550. Email:
journalpermissions@lww.com

Reprints: Information about reprints can be found online at
http://www.lww.com/static/html/reprints.html

AHA Scientific Statement

Standards for Statistical Models Used for Public Reporting of Health Outcomes

An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group

Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council

Endorsed by the American College of Cardiology Foundation

Harlan M. Krumholz, MD, FAHA, Chair; Ralph G. Brindis, MD, MPH; John E. Brush, MD;
David J. Cohen, MD, MSc; Andrew J. Epstein, PhD; Karen Furie, MD, MPH, FAHA;
George Howard, DrPH, FAHA; Eric D. Peterson, MD, MPH; Saif S. Rathore, MS, MPH;
Sidney C. Smith, Jr, MD, FAHA; John A. Spertus, MD;
Yun Wang, PhD; Sharon-Lise T. Normand, PhD

Abstract—With the proliferation of efforts to report publicly the outcomes of healthcare providers and institutions, there is a growing need to define standards for the methods that are being employed. An interdisciplinary writing group identified 7 preferred attributes of statistical models used for publicly reported outcomes. These attributes include (1) clear and explicit definition of an appropriate patient sample, (2) clinical coherence of model variables, (3) sufficiently high-quality and timely data, (4) designation of an appropriate reference time before which covariates are derived and after which outcomes are measured, (5) use of an appropriate outcome and a standardized period of outcome assessment, (6) application of an analytical approach that takes into account the multilevel organization of data, and (7) disclosure of the methods used to compare outcomes, including disclosure of performance of risk-adjustment methodology in derivation and validation samples.

Key Words: AHA Scientific Statements ■ risk adjustment ■ risk assessment ■ quality of health care ■ quality indicators

The American Heart Association convened an interdisciplinary expert writing group to identify the preferred attributes of risk-adjustment models used for public reporting of healthcare providers' outcomes. This statement focuses specifically on the approaches to comparing providers' records on patient outcomes (eg, mortality, hospitalization, quality of life). In the face of expanding efforts to assess providers of cardiovascular health care,[1–5] the attributes identified in this document can serve as a basis for assessing the quality of provider evaluations and promote informative and accurate public reporting of providers' outcomes.

Background

The growing call for accountability in the healthcare system has increased the demand for reports of healthcare providers' performance.[6] Cardiovascular disease has figured prominently in public reporting efforts because of its substantial public health burden, associated costs, and large evidence base supporting management strategies. In the past 2 decades, many groups have developed and publicly reported comparative ratings of providers' cardiovascular healthcare performance.[1–3,7–10] Although some public reporting efforts at the state level have been based on clinical data and have offered operator-specific ratings, at this point most nationwide efforts

have focused on hospitals and have used administrative billing data. Given the potential impact of such ratings in an increasingly competitive healthcare market—and their intent to be viewed as a means of identifying "better" or "worse" providers—a clear need exists for methods that accurately and fairly report providers' outcomes.

Healthcare quality is measured by assessing structures, processes, and/or outcomes.[11] Structure measures, which measure the context in which care is delivered, are relatively straightforward but have only a weak relationship to the care that is delivered and the outcomes that are achieved.[12] Process measures, a characterization of whether the right management strategies are implemented in the appropriate patients at the correct time,[4,5] provide important information about the quality of care. These measures often cover only a narrow range of care for any given patient. Patient outcomes are aggregate markers of quality, integrating structural and process variables that cannot otherwise be measured. Moreover, although structural measures such as procedural volume and process measures such as rates of medication use have been used to characterize the quality of cardiovascular care, outcomes of care have uniquely intuitive appeal to patients.

Outcomes, however, are inherently difficult to measure in ways that allow fair comparisons among providers. First, patients vary in baseline risk for the outcome being studied. For example, Hospital A and Hospital B may have the same mortality rates after bypass surgery, but if Hospital A routinely operates on higher-risk patients, then a reasonable inference is that the quality of care is not the same for these 2 facilities. Second, adverse outcomes may be infrequent and thus highly subject to random variation. Valid comparisons of outcomes therefore require statistical methods that account for important differences in patient characteristics and accurately portray the uncertainty in the estimates. Without these methods, commonly referred to as "risk adjustment," differences in outcomes may only reflect variation in providers' case mix.[13–16] When properly applied, risk-adjustment methods provide a means to determine whether differences in cardiovascular healthcare outcomes across providers may be attributed to providers' behaviors rather than to the patient populations they treat or to the play of chance.

The consequences of inaccurate risk adjustment are notable.[17–20] Poorly performing models can misclassify healthcare providers' outcomes, causing high-quality providers to be incorrectly characterized as having poorer outcomes or, of even more concern, abrogating the shortfalls of lower-quality healthcare providers. Given the central role of risk-adjustment methods in the movement to increase accountability and quality-based purchasing in the healthcare system, there is an urgent need to define the preferred attributes of acceptable comparative risk-adjustment models and to develop criteria against which different risk-adjustment models can be compared.

Attributes of Risk-Adjustment Models

The present interdisciplinary writing group, which consists of individuals with expertise in cardiology, quality of care, outcomes research, statistics, health services research, epidemiology, healthcare policy, and public reporting, identified 7

Preferred Attributes of Models Used for Publicly Reported Outcomes

1. Clear and explicit definition of an appropriate patient sample
2. Clinical coherence of model variables
3. Sufficiently high-quality and timely data
4. Designation of an appropriate reference time before which covariates are derived and after which outcomes are measured
5. Use of an appropriate outcome and a standardized period of outcome assessment
6. Application of an analytical approach that takes into account the multilevel organization of data
7. Disclosure of the methods used to compare outcomes, including disclosure of performance of risk-adjustment methodology in derivation and validation samples

preferred attributes of any risk-adjustment approach used in the comparison of providers' cardiovascular healthcare outcomes. Each of these attributes, by design, is under the control of the organization that is performing the analysis. The attributes include the clear and explicit definition of an appropriate patient sample, clinical coherence of the model variables, sufficiently high-quality and timely data, designation of an appropriate reference time before which covariates are derived and after which outcomes are measured, use of an appropriate outcome and a standardized period of outcome assessment, application of an analytical approach that takes into account the multilevel organization of data, and disclosure of the methods used to compare outcomes, including disclosure of performance of the risk-adjustment methodology in derivation and validation samples (Table).

Several issues deserve attention. First, the attributes described in this scientific statement are intended to be relevant to a wide range of efforts to profile the outcomes of various healthcare providers. The unit that is being analyzed may be a hospital, healthcare system, managed-care organization, physician, group practice, or some other unit that delivers coordinated care. Each of these groups may have special issues with regard to outcomes measurement, but in each case in which the experience of patients is aggregated and reported, the attributes described here have great relevance.

Second, the statement seeks to identify key attributes of models that are used to generate information that is suitable for public reporting. Ideally, all such models would incorporate all of these attributes. All of the attributes should be considered important. The authors recognize that it may not be possible to achieve this ideal in all circumstances. Whether it is acceptable to omit one or more of these attributes may depend on the situation and the alternatives. Thus, the delineation of these attributes also allows for the comparison of proposed models across these domains and for the identification of areas in which the models need improvement.

Third, the issue of administrative data merits special note. The term "administrative data" generally refers to data that are collected for the purposes of resource utilization and cost analyses and includes claims submitted for reimbursement of healthcare services. Databases containing administrative data vary in the scope and detail of the included data. These existing data are often used for profiling efforts because they

are available and often represent the only current national source of information about healthcare outcomes. Administrative claims data, which are known to have variable agreement with information in the medical record, are commonly used to define study samples, derive information to characterize patient risk, and determine outcomes. In this statement, we note issues that are particularly relevant to the use of administrative claims. It is likely that the quality of the claims will improve and, with the advent of electronic health records, more data with clinical information content will be available.

This statement was written to be helpful for the range of data that may be used in profiling efforts. Various data sources have their strengths and weaknesses. Prospectively collected data may be the highest quality, but they are expensive and may include only selected samples of patients because of the need for informed consent. Retrospective medical record review is also expensive, although less so, but it is limited by the quality of medical record documentation. Administrative claims data, which may include information about laboratory values and medications, can be inexpensive but often have been collected for nonclinical purposes and may not accurately or comprehensively represent the patients' conditions. If possible, it should be determined whether the results from models built from administrative claims are good proxies for the results from models based on higher-quality data. If not possible, then other approaches to validation ought to be pursued.

Definition of the Patient Sample

Two samples must be defined in the effort to profile outcomes by healthcare providers: the patients and the providers. For the patients, there should be a clear, reproducible, and appropriate method of identifying the people who should be included in the measurement cohort. This process must balance the interest of including all of the individuals who have the relevant condition ("sensitivity") with the need to avoid including those who do not have the condition ("specificity"). For example, assessments of the outcomes of patients with acute coronary syndromes (ACS) should use valid criteria for the identification of patients while avoiding the inclusion of those without ACS (ie, patients admitted for "rule out ACS" without clear evidence of unstable angina or acute myocardial infarction). As much as possible, available information should be used to confirm the diagnosis and standardize the patient sample across sites. For example, the accuracy of the principal discharge diagnosis code may be improved by including other available information. Patients with a code of acute myocardial infarction could be excluded if their length of stay was <2 days and they did not leave against medical advice because it is unlikely that they were coded correctly.

The provider sample is important in that the groups or individuals that are included, as well as the linkage between providers and patients, must be clearly defined. When profiling hospitals, it is straightforward to assign a patient to an institution unless the patient was transferred into or out of the hospital. In profiling hospitals, the approach to patients who are transferred must be stated and justified. This linkage

between providers and patients is particularly important when physicians or practices are profiled because patients often receive care from a large number of clinicians. Choices about the sample specification need to be clearly stated and justified because they may have important implications for the results of the analysis, particularly with regard to the comparability of institutions.

Clinical Coherence of the Model Variables

Variables in risk-adjustment models should reflect the current understanding of the pathophysiology of cardiovascular disease and the relationship between risk factors and disease outcomes. Thus, clinical judgment and insights from the published literature should guide the selection of candidate variables and the assessment of the model variables. Using clinical judgment and attention to the medical literature to ensure coherent risk adjustment should minimize the influence of variables that may reflect the idiosyncrasies of individual data sets.

The use of variables that convey nonclinical information (eg, race/ethnicity, socioeconomic status) should generally be avoided in these models. The effect of these variables may be mediated through the quality of care, and consequently, adjustment for these factors could confound the results.

Clinical coherence is particularly important for risk-adjustment models that incorporate administrative claims data, which are a commonly used source of data for profiling efforts. Although the ninth revision of the *International Classification of Diseases, Clinical Modification* (ICD-9-CM) contains >15 000 individual codes, many are related, and clinically similar conditions can be distributed across different codes.[21] Grouping ICD-9-CM codes into clinically relevant subgroups makes it more likely that risk-adjustment models will be resistant to minor coding variations while providing full accounting for all of the documented codes associated with the clinical condition.[22]

Data Quality

Efforts should be made to ensure that the data on which the profiling is based are accurate and reliable across the healthcare providers that are being evaluated and, when appropriate, across time. This includes standardizing the definitions of the patient risk variables and where possible disclosing the quality of the data with regard to accuracy and reliability. Such information should be clearly described and substantiated. When data quality is variable, an effort should be made to determine its impact on the profiling results. For example, when administrative data are used, it is important that the results of the analysis be validated against higher-quality clinically derived data because studies indicate that the concordance between administrative claims data and medical record documentation may vary substantially.[23–25]

The data should also be timely. Data that are not recent may not represent the current performance of the group or institution being profiled. In many cases, practical considerations limit the ability to acquire and analyze the data soon after the actual patient encounter. Also, in some cases, data from earlier time periods may be pooled with more recent data to improve the precision of the estimates, which may

compromise the ability to determine whether performance is changing over time. In either case, recent improvements or decrements in performance would be overlooked or diluted. Thus, models could satisfy all of the other criteria and fail to be useful or relevant because of an unacceptable time lag between the performance and when it is reported.

Designation of a Reference Time

Clarity is needed in defining the reference time and ensuring that model variables are measured at or before this time. Events, including complications, that occur after the reference time, or starting time, should not be included as covariates in risk-adjustment models. The goal in using these models is to account for imbalances in the propensity of patients to experience specific outcomes after a specific reference time. Accordingly, models should include information about each patient's characteristics at or before (but not after) that reference time. For example, hemorrhage that occurs after admission, which could be related to the quality of care, should not be counted as a condition that was present on admission when it would appear that the hospital admitted patients with a high degree of comorbid conditions. Exceptions to this general principle may be made for conditions that are detected after a patient has begun treatment but that would, to most clinicians, likely represent a condition that preceded treatment (eg, diagnosis of cancer in a patient hospitalized for acute myocardial infarction).

This issue can be problematic for measurement systems that use administrative claims data.[26,27] The ICD-9-CM codes reported during hospitalizations throughout the United States generally do not differentiate complications that occur during the hospitalization from conditions that are present on admission. Many ICD-9-CM codes represent conditions that might have been present on admission or might have developed during the hospitalization (eg, infection, bleeding, shock), perhaps even as a result of poor-quality care. Models that use patient risk variables that may represent events that occurred as a result of the care provided may inadvertently reward hospitals that poorly manage their patients by allowing them to "adjust" for conditions that arose as a result of deficiencies in care. Consequently, codes that could represent complications should not be included in risk models. Moreover, continuous efforts should be made to improve the administrative claims data so that it would be possible to differentiate complications from comorbidities.

Standardized Assessment of Meaningful Outcomes

First and foremost, the outcome should be meaningful and measured reliably. If the dependent variable cannot reasonably be considered important to patients and related in part to the quality of the clinical care, then it should not be used as an outcome. Any outcome should be measured similarly across providers. When the outcome requires patient response, such as survey information, the response rates must be clear. In the present era, mortality is a common outcome for profiling efforts, but patient-centered outcomes such as health-related quality of life may emerge as these data become more widely measured.

Evaluations of outcomes ideally use a standardized period of end-point evaluation.[28] Nonstandardized periods of assessment, such as the assessment of events during hospitalization, may result in a biased evaluation because healthcare providers have different practice patterns (eg, varying lengths of stay, propensity to transfer patients) that may bias the assessment. If patients are discharged early at some institutions, then the timeframe for the risk of death will be shorter and may make mortality rates appear lower than if a standardized period of follow-up were used. In addition, a substantial number of patients are transferred between healthcare providers during the course of treatment. By focusing on treatment or outcomes provided by only a single provider (eg, treatment at the index admission for patients who are subsequently transferred to another hospital), the period of end-point evaluation is truncated. For example, hospitals that are more likely to transfer patients to skilled nursing facilities or to other acute-care facilities may have a misleadingly low in-hospital mortality rate. Alternatively, hospitals that receive transfer patients who are seriously ill may be inappropriately "penalized" by having higher mortality rates. Previous studies have demonstrated notable variations in hospital performance when comparing risk-adjusted outcomes based on the in-hospital period versus a timeframe based on a standardized end point.[28] Unfortunately, using a standardized period of assessment usually imposes additional requirements for data collection. The resources required to collect these data and ensure that complete follow-up occurs can be considerable. If in-hospital outcomes are used, they should be validated against a standardized period of assessment.

Appropriate Analytical Approach

The appropriate statistical approach is another key feature of these models. Statistical models, particularly those intended for profiling purposes, should account for particular features of the organization of the data.[29] An important aspect of data used for profiling providers relates to the "clustering" of patients within the responsibility of a healthcare provider. For example, in evaluating hospital outcomes, it is the experience of the patients clustered within hospitals that is evaluated. The point is to determine the degree to which the hospital influences the outcomes. Thus, the assumption is that the outcomes of patients within the same hospital may be different than the outcomes of patients in a different hospital. Statistical models used for profiling should take into account that hospital effect, which can be understood as intrahospital correlation.

In addition to the clustered nature of the data, the risk-adjustment model ought to be able to account for the differential amounts of information across providers, measured in terms of the number of observations per provider.[30–32] If 2 hospitals have the same standardized mortality rate, there would be greater confidence in the estimate for the hospital treating 200 cases than for the one with 100 cases. Credible statistical models must be designed to accommodate providers with widely varying sample sizes. Providers with a small number of cases may have observed rates at the extreme ends of the range, but such rates may not accurately reflect their "true" performance. Although some methodologies

specifically exclude providers with fewer cases than a minimum threshold, it is almost impossible to remove all of the variability in sample sizes across providers.

Hierarchical regression models[33] can address the design issues that typically occur in profiling efforts; clustered data and differential information can be addressed by hierarchical generalized linear models.[31,34] These models are used commonly in education, where students are clustered within classrooms and classroom sizes vary.[30] Hierarchical models explicitly quantify interprovider variation and produce better provider-specific estimates for small providers. This particular approach avoids "regression to the mean," a statistical concept describing the tendency for providers who have been identified as outliers in the past to become less extreme in the future.[35] This approach also affords a more realistic assessment of the role of chance in the observed variation between providers.

Public Disclosure

Given the complexities of accurately comparing providers' outcomes, methods used for public reporting should be available in the public domain.[36] The public should have access to a detailed description of the methodological development of the risk-adjustment model and an explicit listing of the variables screened, the manner in which they were measured, and the method of selecting variables for the final model. Information should be available about association of the model variables with the outcome (eg, variable coefficients, odds ratios, confidence intervals) so that the face validity of the results can be evaluated. In addition, the model's performance in derivation and validation sets should be discussed.[18,37] Because of proprietary concerns, it may not be possible for some entities to release all model details (eg, β coefficients and covariance matrices), but at a minimum, measures of the model's performance should be provided, with external auditing for accuracy. The shielding of important information undermines efforts to evaluate the validity of these efforts and confidence in their results.

With regard to these model performance measures, risk-adjustment models should provide measures of discrimination, calibration, and fit. The decision about what constitutes a "good" or "good enough" model will be based more on subjective considerations than on predefined criteria, but users of publicly reported risk-adjustment data should have the benefit of these metrics so that they can make an informed interpretation of presented results. The model performance will depend on the degree to which patient characteristics contribute to the outcome and the availability of variables that reflect factors associated with the outcome. Also, these models should be developed and validated in different samples to assess robustness, and such evaluations should be publicly disclosed. If validation has not been performed, then that should also be reported. In addition, models developed from administrative data should be validated against a model with a more comprehensive description of the patients' clinical conditions, such as medical record data. This validation should include a comparison of how much agreement in classification exists between the administrative and "gold standard" models. In addition to disclosing information about the methods, profiling efforts are needed to provide data on sample size, time interval evaluated, study sample development, data quality, and precision of the estimates.

Conclusions and Recommendations

Interest in the public reporting of outcomes data by providers is growing. Moreover, this information is beginning to have consequence as organizations use it to direct patients to particular institutions and to guide reimbursement strategies. The promise of publicly reporting outcomes ultimately depends on the validity of the measurement system and its transparency. The development of standards is an important step in elevating the performance of current and future measurement systems. In this statement, we proposed 7 attributes of publicly reported risk measures that can be used to evaluate risk-adjustment models. This framework may facilitate the development of models intended for the public reporting of healthcare provider outcomes.

Writing Group Disclosures

Writing Group Disclosures

Writing Group Member	Employment	Research Grant	Other Research Support	Speakers Bureau/ Honoraria	Ownership Interest	Consultant/Advisory Board	Other
Harlan M. Krumholz	Yale University School of Medicine	CV Therapeutics	None	None	None	Advisory board for United Health Care + CV Therapeutics; consultant to Centers for Medicare and Medicaid Services	None
Ralph G. Brindis	Oakland Kaiser Hospital	None	None	None	None	None	None
John E. Brush	Cardiology Consultants, Ltd	None	None	None	None	None	None
David J. Cohen	Beth Israel Deaconess Medical Center	None	None	None	None	None	None
Andrew J. Epstein	Yale University School of Medicine	None	None	None	None	None	None
Karen Furie	Massachusetts General Hospital	None	None	None	None	None	None
George Howard	University of Alabama at Birmingham	None	None	None	None	None	None
Sharon-Lise T. Normand	Harvard University Medical School	PI of contract to monitor and report on all hospital cardiac care quality in the state of Massachusetts	None	None	None	None	Consultant to Yale University/Centers for Medicare and Medicaid Services on public reports of hospital quality*
Eric D. Peterson	Duke University School of Medicine	Schering-Plough; Bristol-Myers Squibb/Sanofi; Merck-Schering	None	None	None	None	None
Saif S. Rathore	Yale University School of Medicine	None	None	None	None	None	None
Sidney C. Smith, Jr	University of North Carolina at Chapel Hill	None	None	None	None	None	None
John A. Spertus	Mid-America Heart Institute	National Institutes of Health and industry support for grants to develop risk-adjusted models of health status outcomes	None	None	Patent pending on translating risk models to patient care	None	None
Yun Wang	Yale–New Haven Health	None	None	None	None	None	None

This table represents the relationships of writing group members that may be perceived as actual or reasonably perceived conflicts of interest as reported on the Disclosure Questionnaire, which all members of the writing group are required to complete and submit. A relationship is considered to be "significant" if the person receives ≥$10 000 during any 12-month period, or ≥5% of the person's gross income; or the person owns ≥5% of the voting stock or share of the entity, or owns ≥$10 000 or more of the fair market value of the entity. A relationship is considered to be "modest" if it is less than "significant" under the preceding definition.

*Modest.

†Significant.

Reviewer Disclosures

Reviewer Disclosures

Reviewer	Employment	Research Grant	Other Research Support	Speakers Bureau/ Honoraria	Ownership Interest	Consultant/ Advisory Board	Other
Ed Havranek	Denver Health and Hospital Authority	None	None	None	None	McKesson	None
Peter Plantes	Veterans Health Administration	None	None	None	None	None	None
Phil Renner	National Committee for Quality Assurance	None	None	None	None	None	None

This table represents the relationships of reviewers that may be perceived as actual or reasonably perceived conflicts of interest as reported on the Reviewer Disclosure Questionnaire, which all reviewers are required to complete and submit.

References

1. Healthgrades. Available at: http://www.healthgrades.com. Accessed April 11, 2005.

2. *U.S. News & World Report*. Top hospitals. Available at: http://www.usnews.com/usnews/health/best-hospitals/tophosp.htm. Accessed September 23, 2005.

3. Solucient 100 Top Hospitals. Available at: http://www.100tophospitals.com. Accessed April 11, 2005.

4. Joint Commission on Accreditation of Healthcare Organizations. A comprehensive review of development and testing for national implementation of hospital core measures. Available at: http://www.jcaho.org/pms/core+measures/cr_hos_cm.htm. Accessed September 23, 2005.

5. Centers for Medicare and Medicaid Services. The specifications manual for national hospital quality measures. Available at: http://www.cms.hhs.gov/quality/hospital/specs.asp. Accessed September 23, 2005.

6. Relman AS. Assessment and accountability: the third revolution in medical care. *N Engl J Med*. 1988;319:1220–1222.

7. California Office of Statewide Health Planning and Development. Healthcare outcomes—clinical data programs. Available at: http://www.oshpd.ca.gov/HQAD/Outcomes/Clinical.htm. Accessed May 2, 2005.

8. New Jersey Department of Health and Senior Services. Cardiac surgery report. Available at: http://www.state.nj.us/health/hcsa/cabmenu.htm. Accessed May 2, 2005.

9. New York State Department of Health. Heart disease. Available at: http://www.health.state.ny.us/nysdoh/heart/heart_disease.htm. Accessed May 2, 2005.

10. Pennsylvania Health Care Cost Containment Council. Pennsylvania's guide to coronary artery bypass graft (CABG) surgery 2003. Available at: http://www.phc4.org/reports/cabg/03/default.htm. Accessed May 2, 2005.

11. Donabedian A. The quality of care. How can it be assessed? *JAMA*. 1988;260:1743–1748.

12. Peterson ED, Coombs LP, DeLong ER, Haan CK, Ferguson TB. Procedural volume as a marker of quality for CABG surgery. *JAMA*. 2004; 291:195–201.

13. Iezzoni LI. Risk adjustment for medical effectiveness research: an overview of conceptual and methodological considerations. *J Investig Med*. 1995;43:136–150.

14. Jencks SF, Daley J, Draper D, Thomas N, Lenhart G, Walker J. Interpreting hospital mortality data: the role of clinical risk adjustment. *JAMA*. 1988;260:3611–3616.

15. Luft HS, Hunt SS. Evaluating individual hospital quality through outcomes statistics. *JAMA*. 1986;255:2780–2784.

16. Halm EA, Chassin MR. Why do hospital death rates vary? *N Engl J Med*. 2001;345:692–694.

17. Iezzoni LI. The risks of risk adjustment. *JAMA*. 1997;278:1600–1607.

18. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med*. 1993;118:201–210.

19. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, Dreyer PI. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72:2155–2168.

20. Shahian DM, Blackstone EH, Edwards FH, Grover FL, Grunkemeier GL, Naftel DC, Nashef SA, Nugent WC, Peterson ED; STS Workforce on Evidence-Based Surgery. Cardiac surgery risk models: a position article. *Ann Thorac Surg*. 2004;78:1868–1877.

21. Department of Health and Human Services. *The International Classification of Diseases, Ninth Revision, 3rd ed, Clinical Modification: ICD-9-CM*. Washington, DC: US Government Printing Office; 1989. DHHS publication PHS 89-1260.

22. Daley J, Iezzoni LI, Shwartz M. *Conceptual and Practical Issues in Developing Risk-Adjustment Methods*. 3rd ed. Chicago, Ill: Health Administration Press; 2003.

23. Jencks SF. Accuracy in recorded diagnoses. *JAMA*. 1992;267: 2238–2239.

24. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. *Ann Intern Med*. 1993;119:844–850.

25. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care*. 2005;43:480–485.

26. McCarthy EP, Iezzoni LI, Davis RB, Palmer RH, Cahalane M, Hamel MB, Mukamal K, Phillips RS, Davies DT Jr. Does clinical evidence support ICD-9-CM diagnosis coding of complications? *Med Care*. 2000; 38:868–876.

27. Lawthers AG, McCarthy EP, Davis RB, Peterson LE, Palmer RH, Iezzoni LI. Identification of in-hospital complications from claims data. Is it valid? *Med Care*. 2000;38:785–795.

28. Jencks SF, Williams DK, Kay TL. Assessing hospital-associated deaths from discharge data: the role of length of stay and comorbidities. *JAMA*. 1988;260:2240–2246.

29. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. *Ann Intern Med*. 2001;135: 112–123.

30. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc*. 1996;159:385–443.

31. Normand ST, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc*. 1997;92:803–814.

32. Normand SL, Zou KH. Sample size considerations in observational health care quality studies. *Stat Med*. 2002;21:331–345.

33. Bryk AS, Raudenbush SW. *Hierarchical Linear Models*. 2nd ed. Thousand Oaks, Calif: Sage Publications; 2002.

34. Daniels MJ, Gatsonis CA. Hierarchical generalized linear models in the analysis of variations in health care utilization. *J Am Stat Assoc*. 1999; 94:29–42.

35. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. New York, NY: John Wiley & Sons; 2004.

36. Iezzoni LI. "Black box" medical information systems: a technology needing assessment. *JAMA*. 1991;265:3006–3007.

37. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130:515–524.