Cardiac Surgery Risk Models: A Position Article

David M. Shahian, MD, Eugene H. Blackstone, MD, Fred H. Edwards, MD, Frederick L. Grover, MD, Gary L. Grunkemeier, PhD, David C. Naftel, PhD, Samer A. M. Nashef, FRCS, William C. Nugent, MD, and Eric D. Peterson, MD, MPH

Lahey Clinic, Burlington, Massachusetts; Cleveland Clinic Foundation, Cleveland, Ohio; University of Florida, Jacksonville, Florida; University of Colorado HSC, Denver, Colorado; Providence Health System, Portland, Oregon; University of Alabama, Birmingham, Alabama; Papworth Hospital, Cambridge, United Kingdom; Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire; Duke Clinical Research Institute, Durham, North Carolina

Differences in medical outcomes may result from disease severity, treatment effectiveness, or chance. Because most outcome studies are observational rather than randomized, risk adjustment is necessary to account for case mix. This has usually been accomplished through the use of standard logistic regression models, although Bayesian models, hierarchical linear models, and machine-learning techniques such as neural networks have also been used. Many factors are essential to insuring the accuracy and usefulness of such models, including selection of an appropriate clinical database, inclusion of critical core variables, precise definitions for predictor variables and endpoints, proper model development, validation, and audit. Risk models may be used to assess the impact of specific predictors on outcome, to aid in patient counseling and treatment selection, to profile provider quality, and to serve as the basis of continuous quality improvement activities.

(Ann Thorac Surg 2004;78:1868–77)

Medical outcome data are often used to compare treatments or providers. Because patient outcomes may be influenced by severity of illness, treatment effectiveness, or chance [1–5], such studies must account for differences in the prevalence of patient risk factors (case mix). In some instances, it is possible to reduce or eliminate outcome variability due to case mix through randomization, which hopefully should balance both known and unknown risk factors. Unfortunately this is impractical in most real-life situations such as the comparison of results among institutions. Other study designs rely on covariate matching or propensity scores [6, 7], techniques which balance only known risk-factors. However, in the majority of existing observational studies in medicine and surgery, *risk adjustment* has been used to account for case mix. Using statistical modeling techniques (typically some form of multivariable regression analysis [8]), investigators study the association between individual risk factors (also referred to as predictor variables or covariates) and outcomes, while holding constant the effect of others. Once the impact of each risk factor is determined from a given population sample, it then becomes possible to estimate the probability of the outcome for patients having particular combinations of these risk factors.

Although researchers have utilized risk models for many years, their broader applicability and critical importance were more fully appreciated after the 1986 release of unadjusted hospital outcome data by the Health Care Financing Administration ([HCFA], now named the Centers for Medicare and Medicaid Services [CMS]). Providers correctly argued that such data did not account for patient severity, and this led directly to the development of a number of high quality clinical databases and risk models, especially in cardiac surgery. Coronary artery bypass grafting (CABG) has been a particular focus of such research, not only because of the desire of cardiac surgeons to improve patient outcomes, but also because regulators and insurers have sought greater control over this high-profile, costly, and frequently performed procedure.

Some of the original cardiac surgery databases were voluntary, such as the Northern New England Cardiovascular Disease Study Group [9], and some were mandated by state or federal law (such as the NY [10], NJ [11], and Veterans Affairs Administration [12] databases). Soon after the HCFA release of unadjusted outcome data in 1986, The Society of Thoracic Surgeons (STS) established an Ad Hoc Committee on Risk Factors for Coronary Artery Bypass Surgery [13]. Subsequently, a committee under the direction of Dr Richard Clark began work on the development of The STS National Cardiac Database (STS NCD). This database was formally established in 1989 and the software was released to STS members in 1990 [14–16]. During the subsequent 13 years, this has evolved to become one of the largest single specialty databases in the world, containing data on more than 2.4 million patients from 60% of United States cardiac programs.

In this review, we present some fundamental aspects of risk model development and validation, the current uses and limitations of such models in cardiac surgery (with special attention to CABG), and prospects for the future. A statistical Appendix is provided to explain less familiar terms and concepts.

Address reprint requests to Dr Shahian, Department of Thoracic and Cardiovascular Surgery, Lahey Clinic, 41 Mall Rd, Burlington, MA 01805; e-mail: david.m.shahian@lahey.org.

Published by Elsevier Inc

0003-4975/04/$30.00
doi:10.1016/j.athoracsur.2004.05.054

Data

Sources

No risk adjustment model is better than the data upon which it is based. Administrative data, such as that from the Centers for Medicare and Medicaid Services MEDPAR database, provide one of the most commonly used sources for observational studies. Such data are readily available, relatively inexpensive, and contain information on millions of patients [2, 17]. However, because these administrative data have been collected primarily for billing purposes rather than for clinical studies, critical variables such as ejection fraction are unavailable, and differentiation of comorbidities from complications is problematic. The latter deficiency may exaggerate the predictive ability of risk models derived from such data by inappropriately including preterminal complications that are highly correlated with mortality. Occasionally this may lead to the paradoxical and incorrect conclusion that such models possess greater predictive accuracy than models derived from clinical databases [2, 17, 18]. Administrative databases may also exclude important variables that are not billable diagnoses ("field saturation") [19]. They also limit the number of secondary diagnoses and generally have insufficient flexibility to properly classify certain comorbidities [17], all of which limit the accuracy of risk models derived from them.

Core Variables

Griffith and associates [20] found that critical clinical variables known to be associated with mortality (eg, ejection fraction, emergency procedures) were not included in the first Pennsylvania CABG database, leading to problematic accuracy. Hannan and associates compared models based on the New York clinical cardiac surgery database (CSRS) with models derived from the New York administrative database (SPARCS) [21] and with the Health Care Financing Administration MEDPAR database [17]. In both instances, models derived from the clinical database were found to provide superior performance. Accuracy of the models based on administrative data were improved substantially by the addition of a few critical clinical variables (ejection fraction, reoperation, left main obstruction).

Tu and associates [4] have determined a limited set of six core variables (age, gender, acuity, reoperation, left main coronary obstruction, ejection fraction) beyond which they believe there is little incremental improvement in model performance. Similarly, Jones and associates [22] from the Cooperative CABG Database Project identified seven core predictor variables (age, gender, left ventricular function, acuity, left main disease, reoperation, number of diseased vessels). They found that acuity, reoperation, and age accounted for the majority of predictive information in most CABG databases. In the STS NCD, 78% of the explained variance from the entire 28 variable model is derived from the eight most important predictors (ie, age, surgical acuity, reoperative status, creatinine level, dialysis, shock, chronic lung disease, and ejection fraction).

All studies confirm that the gold standard for data is a specialty-specific, prospectively maintained clinical database such as the STS NCD. Such databases should contain, at the minimum, a core set of variables that has been demonstrated to be associated with outcome.

Definitions

Certain caveats regarding data apply generally to all regression models including those used for cardiac surgery. One of the most important is strict standardization of definitions, both for predictor variables and for endpoints. Even for a seemingly unambiguous endpoint like mortality, there are important statistical and policy implications of using (1) in-hospital mortality, regardless of when it occurs, (2) 30-day all-cause mortality, regardless of where it occurs, and (3) operative mortality, defined as either (1) or (2). A fixed time period is statistically preferable [1], although more difficult to obtain than in-hospital mortality from databases that are not payer-based. Osswald and colleagues [23, 24] have studied the implications of different definitions of "early" post-CABG mortality, especially in light of advancements in postoperative care. Particularly for higher risk patients in which the early postoperative phase may be prolonged, these investigators assert that the true early mortality will be underestimated unless a complete tally of deaths and their time of occurrence are compiled for the first 6 months after surgery.

Quality

Whenever practical, continuous data should be used as such to avoid the arbitrariness and loss of valuable information that occurs with categorization [25]. However there are some instances in which such categorization may be useful if the goal is to identify a *case*, such as patients with morbid obesity, renal insufficiency, or severe carotid stenosis. This is most applicable when the definition of this categorical state is well-defined and broadly accepted. Transformations of the measurement scale may be required for the values of the variables to be commensurate with the assumptions of the risk factor model being used (eg, so-called linearizing transformations). Data entry software should contain internal quality controls for out of range, inconsistent, contradictory, and missing data. Ideally, values for missing data should be substituted using multiple imputation techniques [26, 27]. Institutions should receive periodic reports regarding their data quality including any anomalies in their data recording compared with regional and national averages. All these features are included in the STS NCD. Particularly in situations where risk-adjusted outcomes are used to assess provider performance, there should be regular independent auditing of the data to assure accuracy and completeness.

Model Development

Risk model development requires considerable statistical expertise and judgment, a caveat that is sometimes forgotten in this era of ubiquitous, powerful, off-the shelf statistical software. For example, the type of modeling strategy and validation techniques may differ depending on whether the purpose of the model is description of relationships (ie, comparison of providers or treatments) or prediction of future events [28, 29]. Spiegelhalter [29] has demonstrated the importance of such considerations when the aim of the model is probabilistic prediction to aid in patient selection and counseling. Even when the same basic model is used, Naftel [30] demonstrated numerous technical reasons that different multivariable equations may be developed by different statisticians from the same data.

Three principal techniques have been utilized for the construction of cardiac surgery risk models. Bayesian models were used initially for the STS NCD because they are robust with regard to missing data, an important problem in the early database experience. As data completeness improved, logistic models were substituted in 1995 [31, 32]. Logistic regression models continue to be the most common statistical technique for cardiac surgery risk modeling, not only for the STS NCD but also for those developed in New York [33], the Veterans Affairs Administration [34], and the Northern New England Cardiovascular Disease Study Group [35]. Some groups have used simple, additive scores with weights derived from the logistic regression model [11, 36]. Comparative studies [37] have generally demonstrated that logistic models offer the best overall performance.

Some have suggested that the next major advance in model performance would come from the application of algorithmic models, sometimes called machine-learning techniques [38], of which artificial neural networks are an example [39]. These models permit complex, nonlinear information processing, thus avoiding one of the constraints of logistic models. However, two studies [40, 41], one using 80,606 patients from the STS NCD [40], failed to demonstrate any significant improvement over logistic or Bayesian models. Potential disadvantages of machine-learning techniques include the large amount of data required for model development, and the tendency of some methods to overfit the data and others to perform too conservatively. Furthermore, just as with traditional statistical models, algorithmic techniques generally model past data better than they predict future events.

Strategies for logistic regression model development include data reduction techniques and variable selection, interaction terms and transformations of variables where appropriate, imputation of missing data, verification of model assumptions, and model validation [8, 42, 43]. One fundamental and still controversial question is the number of variables to include in a risk model. An excessive number of covariates may lead to some predictors having statistical significance but not biological or clinical relevance, over-fitting of the model, numerical instability, and increased cost and difficulty of data collection [4, 8, 42, 43]. Harrell and colleagues [8, 42] have recommended that the number of covariates considered for inclusion in such models be less than one-tenth the number of end points observed in the data set (percent occurrence, x sample size, for early events), which typically requires some data reduction technique to decrease the number of candidate variables. Univariate screening of candidate covariates followed by forward or backward stepwise selection is commonly used for this purpose. However some statisticians including Harrell [8] have criticized this variable selection technique on theoretical grounds, and research continues on the relative merits of parsimonious models versus those that retain most or all potential predictors [29, 44, 45]. The use of machine-learning variable selection methods such as *bagging* (bootstrap aggregating) is also gaining popularity [46, 47].

Finally, most studies demonstrate that models have inferior performance when applied to patient groups other than the one from which they were developed. Ivanov and colleagues [48] found that institution or region-specific custom models performed best, followed by recalibrated models using the covariate set from a ready-made model, but with different weights. Applying unmodified off-the-shelf models to other groups of patients provided the least satisfactory performance. In contrast, Nashef and colleagues [49] recently reported that the additive EuroSCORE cardiac surgery risk-prediction model functioned well when applied to North American populations.

Model Validity

Any statistical risk model must be scrutinized to determine whether it functions reliably for its intended purpose. Numerous types of validity have been summarized by Daley [1] including face validity (the model is reasonable to experts), content validity (all important variables have been included), attributional validity (risk adjustment is adequate to insure that differences in outcome are not due to patient characteristics), and predictive validity.

The predictive validity of a model is a measure of how well it performs on a data set other than the one from which it was developed. This test data may be internal or external. In the former, only the original data set is used. Numerous techniques are available to segment the original data including simple data splitting (the whole data set is randomly and only once split into development and validation or test subsets) and more sophisticated cross-validation techniques (eg, leave-one-out cross validation and *K*-fold cross validation) that use repeated resampling from the original data set [8, 50]. Harrell [8] has recommended using the entire data set for model development, then validating by using another technique such as bootstrapping. External validation uses a completely new data set to validate the model, perhaps originating from a different state, country, or hospital.

Regardless of whether internal or external test data are used, certain tests are commonly used to assess how well the model fits the data. Two of the most common are calibration (reliability) and discrimination (resolution). Calibration assesses the extent to which the model assigns appropriate risk to the population under consideration. It answers the question: in 100 patients with the same estimated mortality risk (R%) as mine, would the observed number of deaths equal R? Calibration may be measured by a number of techniques including the Pearson χ^2 statistic, the unweighted residual sum-of-squares, or the Murphy decomposition of the Brier mean probability score [43, 51]. The most commonly used measure of calibration is the degree of concordance between deciles of observed and expected risk, the Hosmer-Lemeshow test [43]. Most CABG risk models are well calibrated overall but may produce estimates that are too extreme in the lowest and especially the highest risk subsets of patients, particularly in small data sets [45, 52]. Shrinkage of regression coefficients may provide more accurate and realistic prediction for future patients [8, 45, 52, 53].

Discrimination is the more demanding test and measures the tradeoff between the specificity and sensitivity of the risk model at various probability cut points (ie, what probability do you require to assign the patient to a particular outcome category?). It asks the question: how well does the model separate patients who die from those who survive? This may be interpreted as the percentage of discordant (meaning one death and one survivor) patient-pairs for which the model predicts a higher mortality risk for the patient who actually dies [43, 54, 55]. This percentage is equal to the area under the receiver operating characteristic curve (ROC) [43, 55] and is called the c-index or c-statistic [8, 42]. Unfortunately, most CABG risk models have only moderate discrimination as measured by the c-index. This has significant implications for their use in individual patient counseling [54, 56]. Risk models may accurately predict that 3 of 100 patients with a given set of risk factors will die postoperatively, but they cannot identify which 3.

The similar and limited discrimination of most CABG risk models (VA, Northern New England Cardiovascular Disease Study Group, STS, NY) suggests that we still await the quantum improvement in risk prediction anticipated by Steen [39] more than a decade ago. The performance of current models is limited by as yet unknown predictors, difficulty in measuring or representing certain complex clinical states, random catastrophic events, such as a serious protamine reaction or sudden hemorrhage, and similar occurrences that are rare in the population but important in the individual patient [22, 40, 57–59]. Because known patient characteristics will never explain all the variances in cardiac surgery outcomes, the performance of any risk model has inherent limitations.

Uses of Risk Models

One of the most important uses of cardiac risk models, including the STS risk model, has been for academic research. Typically this has involved estimation of the effect of risk factors or particular therapies on patient outcome. Logistic models are well suited to this function because they readily provide odds ratios for each risk factor. Studies of preoperative risk factors derived from the STS NCD include the impact of race [60], gender [61], and obesity [62]. Similar investigations of therapeutic options have included the value of IMA use [63, 64], β blockade [65], and off-pump techniques [66]. The STS NCD has also been utilized to clarify our understanding of the relationship between volume and outcome for CABG [67–69].

A second use for risk models is the development of tools that aid in everyday patient management. These would include patient care algorithms or critical pathways scientifically based on risk-adjusted studies [58]. Risk models may be used to facilitate individual patient counseling or as a decision support tool for clinicians choosing between different interventions (eg, coronary artery bypass vs percutaneous angioplasty). Pocket cards (Northern New England Cardiovascular Disease Study Group) [70] or handheld computers [71] have been used to generate bedside risk estimates for individual patients. Obviously, in order to properly apply such methods, values for each risk factor in the model must be available. Because of the modest discrimination of most risk models, which limits the accuracy of individual patient prediction, it is recommended that such information not be presented to patients simply as probability point estimates. Rather, these estimates should be accompanied by confidence limits that demonstrate their uncertainty.

One of the most common uses of risk models is to compare provider performance. This is statistically challenging from the outset because of the low incidence of the binary outcome (operative mortality) as well as the highly variable and often small sample sizes from different providers. Typically, as in New York State, profiling has been achieved by aggregating the probabilities of death of each patient treated by a given provider based on the results of logistic modeling. This aggregate predicted mortality is used together with the observed provider mortality to construct a ratio of observed to expected mortalities (O/E). However, modern research in provider profiling has demonstrated potential deficiencies with this approach. Standard techniques may not accurately reflect the unobserved true mortality of low volume providers, and they do not adequately account for clustering (nonrandom allocation) of observations within providers (such as the prevalence of heart failure patients at transplant centers [58]). The net effect may be an underestimation of random interprovider variability, an overestimation of systematic interprovider variability, and an increased likelihood of falsely classifying a provider as an outlier. Hierarchical or multilevel models have been designed to address these concerns and some advocate their use for profiling whenever feasible [44, 72–77].

Whatever method is used for provider profiling, identification of statistical outliers is only a starting point for further analysis [58, 78]. All risk models have inherent limitations, and the results obtained from different models and by different statisticians may vary [28–30, 52, 74].

The results derived from any risk model should be regarded as one element to be considered in conjunction with other traditional indices of competent surgical care. The Society of Thoracic Surgeons firmly holds that those surgical programs identified as statistical outliers should not be arbitrarily labeled as substandard. The coding practices of outlying providers and each individual mortality should be carefully analyzed, and structural causes should be sought to explain any truly aberrant results [12, 79–82].

Finally, risk-adjusted outcomes have also been used in northern New England [83] and Minnesota [84] as the basis for confidential continuous quality improvement activities, and in the Veterans Affairs Administration for both confidential monitoring of performance and continuous quality improvement [12, 80]. Here the main goal is not public accountability but rather provider-initiated determination of best practice, benchmarking, and regional or system-wide improvement. This has resulted in mortality reduction that appears comparable with that achieved using public report cards [85]. The Society of Thoracic has implemented process improvement initiatives based on analyses derived from the STS NCD [64].

Limitations and Disadvantages of Risk Models

As previously described, risk models do not have perfect discrimination, the ability to predict the death of specific individuals. Because our ability to determine expected outcome is limited, risk-adjusted mortality estimates derived from the observed to expected mortality ratio are also subject to error. Publication of risk-adjusted mortality "accurate" to the nearest hundredth of a percentage may be misleading to the public, and all the more so when these are not accompanied by confidence limits. Furthermore, even the best available risk models explain only a small proportion of the variability in cardiac surgery outcomes, a significant liability if the goal is to assess interprovider differences in quality of care based on their relative risk-adjusted mortalities [28, 86]. Statisticians continue to evolve new and more sophisticated techniques to model complex biological phenomena, and it is hopeful that the performance of risk models will continue to improve.

From a health policy perspective, some argue that current risk models place too much emphasis on mortality as the sole endpoint, which may ultimately decrease access to surgery for those who might benefit most (high-risk case avoidance) [57, 72, 87, 88]. For similar reasons, such models may encourage gaming of the reporting system when used for provider profiling [72]. Emphasis on outcome endpoints such as mortality may deflect attention from important process and structural aspects of care. Finally, challenges common to all cardiac surgery databases include their cost (especially at a time when revenues from heart surgery are declining) and privacy issues related to the HIPAA.

Current Status and Future Direction

Between 1990 and 1997, the number of centers contributing patients to the STS NCD grew from 105 to 450, and the database currently contains clinical data on 2.4 million patients [16, 78, 89]. Results during the decade from 1990 through 1999 demonstrate a progressive increase in preoperative risk and a decline in both observed mortality and the observed to expected mortality ratio [89].

Although it is already the dominant cardiac surgical database in the world, it is the goal of the STS to achieve 100% participation of all cardiac surgery providers in the United States. Along with increased efforts to validate the accuracy of the submitted data, this will eliminate any remaining concerns about the voluntary nature of the database, although generally there has been a close correlation between the results obtained from the voluntary databases (STS NCD, Northern New England Cardiovascular Disease Study Group) and mandatory databases (VA, NY) [78, 89]. Recent studies matching STS NCD results with those from the Centers for Medicare and Medicaid Services claims data suggest substantial agreement in both completeness and observed mortality. The STS NCD should become the dominant source of accurate information for the federal and state governments regarding outcome and reimbursement issues, and there is also the opportunity to partner with the industry in assessing new technologies.

A systematic review of the STS NCD was begun in 1997 [16, 78]. The STS Definitions Committee worked with the American College of Cardiology to eliminate unimportant variables and to develop a new data dictionary. The original 512 fields were reduced to 217 core fields and 255 extended fields. Discussions continue among representatives of the major cardiac surgery and cardiology databases to resolve inconsistent data definitions.

The Duke Clinical Research Institute became the data analysis and warehouse center for the STS NCD beginning in 1998. In addition to their sophisticated resources for data cleaning and verification, the Duke Clinical Research Institute provides detailed semiannual reports comparing individual programs with their region and the overall STS national data set. Regular national meetings of hospital data managers have substantially enhanced uniformity of coding practices and completeness.

From models that focused primarily on CABG mortality, the STS NCD has evolved to a family of related risk models for prediction of outcomes after valve replacement (with or without CABG) [90, 91], congenital heart surgery [92], and general thoracic surgery. There is an increased awareness of the importance of endpoints other than operative mortality, including perioperative morbidity and length of stay [15, 16, 78, 93, 94]. Postoperative complications and readmissions, which are delayed complications, appear to measure different and complementary aspects of care compared with hospital mortality [95, 96], and not all measures of outcome are equally indicative of program quality. Future emphasis will also include documentation of long-term follow-up data, functional status and quality of life measures, and the relationship of clinical factors to hospital costs. Finally, there will be an increasing emphasis on process measures (eg, internal mammary artery and β blocker use) to assess and improve provider performance [58, 64, 78]. This approach may offer the greatest potential to enhance the results of all cardiac surgery programs and to reduce interprovider variability. It also diminishes our reliance on sophisticated, yet still imperfect, methods of risk-adjustment.

Cardiac surgery remains at the forefront of risk model development and clinical quality monitoring. With advancements in statistical methodology, expanding enrollment in major databases such as the STS NCD and European databases, and with the firm commitment of cardiac surgeons, our profession will maintain its leadership in this vital area of health care.

References

1. Daley J. Criteria by which to evaluate risk-adjusted outcomes programs in cardiac surgery. Ann Thorac Surg 1994;58:1827–35.
2. Iezzoni LI. The risks of risk adjustment. JAMA 1997;278: 1600–7.
3. Iezzoni LI. Risk adjustment for measuring healthcare outcomes. Chicago: Health Administration Press, 1997.
4. Tu JV, Sykora K, Naylor CD. Assessing the outcomes of coronary artery bypass graft surgery: how many risk factors are enough? Steering Committee of the Cardiac Care Network of Ontario. J Am Coll Cardiol 1997;30:1317–23.
5. Luft HS, Romano PS. Chance, continuity, and change in hospital mortality rates. Coronary artery bypass graft patients in California hospitals, 1983 to 1989. JAMA 1993;270: 331–7.
6. Grunkemeier GL, Payne N, Jin R, Handy JR Jr. Propensity score analysis of stroke after off-pump coronary artery bypass grafting. Ann Thorac Surg 2002;74:301–5.
7. Blackstone EH. Comparing apples and oranges. J Thorac Cardiovasc Surg 2002;123:8–15.
8. Harrell FE Jr. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag, 2001.
9. O'Connor GT, Plume SK, Olmstead EM, et al. A regional prospective study of in-hospital mortality associated with coronary artery bypass grafting. The Northern New England Cardiovascular Disease Study Group. JAMA 1991; 266:803–9.
10. Hannan EL, Kilburn H Jr, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. JAMA 1994;271:761–6.
11. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation 1989;79:I3–12.
12. Hammermeister KE, Johnson R, Marshall G, Grover FL. Continuous assessment and improvement in quality of care. A model from the Department of Veterans Affairs cardiac surgery. Ann Surg 1994;219:281–90.
13. Kouchoukos NT, Ebert PA, Grover FL, Lindesmith GG. Report of the Ad Hoc Committee on Risk Factors for Coronary Artery Bypass Surgery. Ann Thorac Surg 1988;45: 348–9.
14. Clark RE. The development of The Society of Thoracic Surgeons voluntary national database system: genesis, issues, growth, and status. Best Pract Benchmarking Healthc 1996;1:62–9.
15. Edwards FH. Evolution of The Society of Thoracic Surgeons National Cardiac Surgery Database. J Invasive Cardiol 1998;10:485–8.
16. Ferguson TB Jr, Dziuban SW Jr, Edwards FH, et al. The STS National Database: current changes and challenges for the new millennium. Committee to Establish a National Database in Cardiothoracic Surgery, The Society of Thoracic Surgeons. Ann Thorac Surg 2000;69:680–91.
17. Hannan EL, Racz MJ, Jollis JG, Peterson ED. Using Medicare claims data to assess provider quality for CABG surgery: Does it work well enough? Health Serv Res 1997; 31:659–78.
18. Landon B, Iezzoni LI, Ash AS, et al. Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. Inquiry 1996;33:155–66.
19. Finlayson EV, Birkmeyer JD, Stukel TA, Siewers AE, Lucas FL, Wennberg DE. Adjusting surgical mortality rates for patient comorbidities: More harm than good? Surgery 2002;132:787–94.
20. Griffith BP, Hattler BG, Hardesty RL, Kormos RL, Pham SM, Bahnson HT. The need for accurate risk-adjusted measures of outcome in surgery. Lessons learned through coronary artery bypass. Ann Surg 1995;222:593–8.
21. Hannan EL, Kilburn H Jr, Lindsey ML, Lewis R. Clinical versus administrative data bases for CABG surgery: Does it matter? Med Care 1992;30:892–907.
22. Jones RH, Hannan EL, Hammermeister KE, et al. Identification of preoperative variables needed for risk adjustment of short-term mortality after coronary artery bypass graft surgery. The Working Group Panel on the Cooperative CABG Database Project J Am Coll Cardiol 1996;28:1478–87.
23. Osswald BR, Tochtermann U, Schweiger P, et al. Minimal early mortality in CABG: Simply a question of surgical quality? Thorac Cardiovasc Surg 2002;50:276–80.
24. Osswald BR, Blackstone EH, Tochtermann U, Thomas G, Vahl CF, Hagl S. The meaning of early mortality after CABG. Eur J Cardiothorac Surg 1999;15:401–7.
25. Altman DG. Categorizing continuous variables. Br J Cancer 1991;64:975.
26. Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken: Wiley-Interscience, 2002.
27. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res 1999;8:3–15.
28. Krumholz HM. Mathematical models and the assessment of performance in cardiology. Circulation 1999;99:2067–9.
29. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.
30. Naftel DC. Do different investigators sometimes produce different multivariable equations from the same data? J Thorac Cardiovasc Surg 1994;107:1528–9.
31. Edwards FH, Grover FL, Shroyer AL, Schwartz M, Bero J. The Society of Thoracic Surgeons National Cardiac Surgery Database: current risk assessment. Ann Thorac Surg 1997; 63:903–8.
32. Shroyer AL, Plomondon ME, Grover FL, Edwards FH. The 1996 coronary artery bypass risk model: The Society of Thoracic Surgeons Adult Cardiac National Database. Ann Thorac Surg 1999;67:1205–8.
33. Hannan EL, Kilburn H Jr, O'Donnell JF, Lukacik G, Shields EP. Adult open heart surgery in New York State. An analysis of risk factors and hospital mortality rates. JAMA 1990;264:2768–74.
34. Grover FL, Shroyer AL, Hammermeister KE. Calculating risk and outcome: the Veterans Affairs database. Ann Thorac Surg 1996;62:S6–11.
35. O'Connor GT, Plume SK, Olmstead EM, et al. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery. Northern New England Cardiovascular Disease Study Group. Circulation 1992;85: 2110–8.
36. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk

evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999;16:9–13.

37. Marshall G, Grover FL, Henderson WG, Hammermeister KE. Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery. Stat Med 1994;13:1501–11.
38. Breiman L. Statistical modeling: the two cultures. Statistical Science 2001;16:199–231.
39. Steen PM. Approaches to predictive modeling. Ann Thorac Surg 1994;58:1836–40.
40. Lippmann RP, Shahian DM. Coronary artery bypass risk prediction using neural networks. Ann Thorac Surg 1997; 63:1635–43.
41. Orr RK. Use of a probabilistic neural network to estimate the risk of mortality after cardiac surgery. Med Decis Making 1997;17:178–85.
42. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
43. Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 1989.
44. Grunkemeier GL, Zerr KJ, Jin R. Cardiac surgery report cards: making the grade. Ann Thorac Surg 2001;72:1845–8.
45. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modeling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19:1059–79.
46. Blackstone EH. Breaking down barriers: helpful breakthrough statistical methods you need to understand better. J Thorac Cardiovasc Surg 2001;122:430–9.
47. Breiman L. Bagging predictors. Machine Learn 1996;24: 123–40.
48. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation 1999;99:2098–104.
49. Nashef SA, Roques F, Hammill BG, et al. Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American cardiac surgery. Eur J Cardiothorac Surg 2002;22:101–5.
50. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall, 1993.
51. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5:421–33.
52. Steyerberg EW, Ivanov J, Tu JV, Naylor CD, Krumholz HM. Ranking of surgical performance. Circulation 2000;102:E61–62.
53. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9:1303–25.
54. Grunkemeier GL, Zerr KJ, Jin R. Reply. Ann Thorac Surg 2002;74:1749.
55. Grunkemeier GL, Jin R. Receiver operating characteristic curve analysis of clinical risk models. Ann Thorac Surg 2001;72:323–6.
56. Pinna-Pintor P, Bobbio M, Colangelo S, et al. Inaccuracy of four coronary surgery risk-adjusted models to predict mortality in individual patients. Eur J Cardiothorac Surg 2002; 21:199–204.
57. Parsonnet V. Risk stratification in cardiac surgery: Is it worthwhile? J Card Surg 1995;10:690–8.
58. Grover FL, Hammermeister KE, Shroyer AL. Quality initiatives and the power of the database: what they are and how they run. Ann Thorac Surg 1995;60:1514–21.
59. Warner BA. Thoughts and considerations on modeling coronary bypass surgery risk. Ann Thorac Surg 1997;63: 1529–30.
60. Hartz RS, Rao AV, Plomondon ME, Grover FL, Shroyer AL. Effects of race, with or without gender, on operative mortality after coronary artery bypass grafting: a study using The Society of Thoracic Surgeons national database. Ann Thorac Surg 2001;71:512–20.
61. Edwards FH, Carey JS, Grover FL, Bero JW, Hartz RS. Impact of gender on coronary bypass operative mortality. Ann Thorac Surg 1998;66:125–31.
62. Prabhakar G, Haan CK, Peterson ED, Coombs LP, Cruzzavala JL, Murray GF. The risks of moderate and extreme obesity for coronary artery bypass grafting outcomes: a study from The Society of Thoracic Surgeons' database. Ann Thorac Surg 2002;74:1125–31.
63. Edwards FH, Clark RE, Schwartz M. Impact of internal mammary artery conduits on operative mortality in coronary revascularization. Ann Thorac Surg 1994;57:27–32.
64. Ferguson TB Jr, Peterson ED, Coombs LP, et al. Use of continuous quality improvement to increase use of process measures in patients undergoing coronary artery bypass graft surgery: a randomized controlled trial. JAMA 2003; 290:49–56.
65. Ferguson TB Jr, Coombs LP, Peterson ED. Preoperative beta-blocker use and mortality and morbidity following CABG surgery in North America. JAMA 2002;287:2221–7.
66. Cleveland JC Jr, Shroyer AL, Chen AY, Peterson E, Grover FL. Off-pump coronary artery bypass grafting decreases risk-adjusted mortality and morbidity. Ann Thorac Surg 2001;72:1282–8.
67. Crawford FA Jr, Anderson RP, Clark RE, et al. Volume requirements for cardiac surgery credentialing: a critical examination. The Ad Hoc Committee on Cardiac Surgery Credentialing of The Society of Thoracic Surgeons. Ann Thorac Surg 1996;61:12–6.
68. Clark RE. Outcome as a function of annual coronary artery bypass graft volume. The Ad Hoc Committee on Cardiac Surgery Credentialing of The Society of Thoracic Surgeons. Ann Thorac Surg 1996;61:21–6.
69. 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.
70. Eagle KA, Guyton RA, Davidoff R, et al. ACC/AHA Guidelines for Coronary Artery Bypass Graft Surgery: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Revise the 1991 Guidelines for Coronary Artery Bypass Graft Surgery). American College of Cardiology/ American Heart Association. J Am Coll Cardiol 1999;34: 1262–347.
71. Bernstein AD, Parsonnet V. Bedside estimation of risk as an aid for decision-making in cardiac surgery. Ann Thorac Surg 2000;69:823–8.
72. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. Ann Thorac Surg 2001;72:2155–68.
73. Normand SLT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. J Am Stat Assoc 1997;92:803–14.
74. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc (Series A) 1996;159:385–443.
75. Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. Stat Med 1994;13:889–903.
76. DeLong E. Hierarchical modeling: its time has come. Am Heart J 2003;145:16–8.
77. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. Stat Med 2002;21:3291–315.
78. Grover FL, Shroyer AL, Hammermeister K, et al. A decade's experience with quality improvement in cardiac surgery using the Veterans Affairs and The Society of Thoracic Surgeons national databases. Ann Surg 2001;234: 464–72.
79. Grover FL, Johnson RR, Shroyer AL, Marshall G, Hammermeister KE. The Veterans Affairs Continuous Improvement in Cardiac Surgery Study. Ann Thorac Surg 1994;58:1845–51.
80. Hammermeister KE. Participatory continuous improvement. Ann Thorac Surg 1994;58:1815–21.

81. Daley J, Forbes MG, Young GJ, et al. Validating risk-adjusted surgical outcomes: site visit assessment of process and structure. National VA Surgical Risk Study. J Am Coll Surg 1997;185:341–51.
82. Dziuban SW Jr. Using information from databases to improve clinical practice: lessons learned under fire. Ann Thorac Surg 1997;64:S64–7.
83. O'Connor GT, Plume SK, Olmstead EM, et al. A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. The Northern New England Cardiovascular Disease Study Group. JAMA 1996;275:841–6.
84. Arom KV, Petersen RJ, Orszulak TA, et al. Establishing and using a local/regional cardiac surgery database. Ann Thorac Surg 1997;64:1245–9.
85. Peterson ED, Delong ER, Jollis JG, Muhlbaier LH, Mark DB. The effects of New York's bypass surgery provider profiling on access to care and patient outcomes in the elderly. J Am Coll Cardiol 1998;32:993–9.
86. Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York State's approach. N Engl J Med 1995; 332:1229–33.
87. Report of the Ad Hoc Committee on Physician-Specific Mortality Rates for Cardiac Surgery. Ann Thorac Surg 1993;56:1200–2.
88. Jones RH. In search of the optimal surgical mortality. Circulation 1989;79:I132–6.
89. Ferguson TB Jr, Hammill BG, Peterson ED, DeLong ER, Grover FL. A decade of change–risk profiles and outcomes for isolated coronary artery bypass grafting procedures, 1990–1999: a report from the STS National Database Committee and the Duke Clinical Research Institute. Society of Thoracic Surgeons. Ann Thorac Surg 2002;73:480–9.
90. Edwards FH, Peterson ED, Coombs LP, et al. Prediction of operative mortality after valve replacement surgery. J Am Coll Cardiol 2001;37:885–92.
91. Jamieson WR, Edwards FH, Schwartz M, Bero JW, Clark RE, Grover FL. Risk stratification for cardiac valve replacement. National Cardiac Surgery Database. Database Committee of The Society of Thoracic Surgeons. Ann Thorac Surg 1999;67:943–51.
92. Mavroudis C, Gevitz M, Ring WS, McIntosh CL, Schwartz M. The Society of Thoracic Surgeons National Congenital Heart Surgery Database Report: analysis of the first harvest (1994–1997). Ann Thorac Surg 1999;68:601–24.
93. Peterson ED, Coombs LP, Ferguson TB, et al. Hospital variability in length of stay after coronary artery bypass surgery: results from The Society of Thoracic Surgeon's National Cardiac Database. Ann Thorac Surg 2002;74:464–73.
94. Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. Ann Thorac Surg 2003;75:1856–65.
95. Hannan EL, Racz MJ, Walford G, et al. Predictors of readmission for complications of coronary artery bypass graft surgery. JAMA 2003;290:773–80.
96. Silber JH, Rosenbaum PR, Schwartz JS, Ross RN, Williams SV. Evaluation of the complication rate as a measure of quality of care in coronary artery bypass graft surgery. JAMA 1995;274:317–23.
97. Kreft I, DeLeeuw J. Introducing multilevel modeling. London: Sage Publications, 1998.
98. Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modeling. London: Sage Publications, LTD; 1999.
99. Anderson RP, Jin R, Grunkemeier GL. Understanding logistic regression analysis in clinical reports: an introduction. Ann Thorac Surg 2003;75:753–7.
100. Armitage P, Berry G. Statistical methods in medical research. Oxford: Blackwell Scientific Publications, 1987.

Appendix

Bayesian Modeling

In contrast to frequentist statistics, Bayesian models treat population parameters as random quantities with probability distributions, not fixed points. Bayesian models require the specification of a prior probability distribution, which may, however, be vague or noninformative. When combined with the observed data, a new revised estimate of the population parameter is determined, known as the posterior probability. The more observed data are available, the less the impact of the prior probability, and visa versa. Frequentist statisticians argue that the incorporation of Bayesian prior probabilities is too subjective, whereas Bayesians counter that this subjectivity is also present in traditional statistical models but not explicitly acknowledged. Perhaps the most serious concern regarding Bayesian models such as those used in early iterations of the STS NCD was failure to adequately account for correlation among variables (redundancy, covariance).

Bootstrap Technique

Efron introduced the bootstrap technique in 1979 as a general method for estimating standard errors and confidence intervals of any test statistic [50]. This method relies on repeated bootstrap samples, each of which involves re-sampling with replacement from the original data set. All samples will have the same number of observations as the original, but in each bootstrap sample any of the original observations may be included once, more than once, or not at all. The standard deviation of the bootstrap replicates is taken as the standard error of the data. The bootstrap technique has found many other applications, including its use in risk factor identification [47]. In this technique, known as bagging or bootstrap aggregation, analyses are performed on all bootstrap data sets and the results are aggregated.

Brier Mean Probability Score

The Brier mean probability score was a formula described by Brier in 1950 to assess the predictive accuracy of weather forecasting, but it is equally applicable to any situation with dichotomous outcomes such as postoperative mortality [29]. The Brier score varies between 0 and 1, with smaller values representing better performance and a score of 0 indicating perfect prediction. The Brier score may be decomposed into components that assess different aspects of predictive accuracy, including calibration and resolution.

Hierarchical Models

Many real-life situations are inherently multilevel or hierarchical [73–75, 77, 97, 98]. In medical studies, patients are grouped by physicians and physicians are grouped within hospitals. In a two-level model, the first or lower level (eg, patient) is sometimes referred to as the micro level, and the second or higher level (eg, surgeon or hospital) is the macro level, group, context, or cluster.

Often there is interest in understanding the interaction between the several levels of such a hierarchy, or in evaluating macro-level performance based on micro-level data. For example, in provider profiling, quality at the group level (physician or hospital) has often been assessed through the use of aggregated, risk-adjusted patient outcomes. However, such traditional approaches ignore an important aspect of multilevel structure, namely that subjects within a group or cluster are more similar to each other than they are to subjects in other clusters. For example, patients treated at a heart failure center are more similar to each other than they are to patients at an institution without this special interest. Such clustering may also exist when longitudinal studies are conducted using repeated observations from the same subjects. In either situation, intraclass correlation or clustering effectively reduces the amount of independent data available, increases the standard errors of estimates, and if not accounted for may exaggerate the significance of performance differences between groups.

Through the use of hierarchical models (also called random effects models, mixed models, or random coefficient regression models), the amount of random and systematic variation between groups can be more accurately partitioned, more appropriate and conservative estimates of between-group variability are obtained, and the impact of specific macro-level characteristics can be better assessed. Furthermore, these models improve the estimates from groups with relatively smaller numbers of observations by shrinking their results closer to the mean of the remaining groups.

Jackknife

From a data set with *n* data points, *n*-jackknife samples are derived, each of which has exactly one observation deleted. From these, estimates of biases and standard errors can be obtained. The jackknife was the predecessor of the bootstrap. It is computationally simpler and may be thought of as an approximation to the bootstrap [50].

K-fold Cross Validation

K-fold cross validation is similar to leave-one-out cross validation. The original data are randomly divided into *k* groups of equal or roughly equal size. A model is developed from *k*-1 groups and its accuracy is assessed using the excluded group. The process is repeated *k* times and the predictive accuracy is averaged [8, 50].

Leave-One-Out Cross Validation

For a data set with *n* observations, training is conducted on *n* different subsets of data, each of which has one data point left out. Each excluded observation is predicted by the model obtained from the remaining data points, and the average predictive accuracy of the models is determined [8, 50].

Logistic Regression Model

The logistic regression model is a type of multivariable model in which, unlike standard regression analysis, the dependent or outcome variable is qualitative [43, 99]. In medical applications the most common example is binary logistic regression to estimate the probability of a dichotomous outcome such as mortality. Given a particular combination of predictor variables, the probability of event occurrence is constrained by the logistic equation to the range of 0 to 1. Rearrangement of the logistic equation demonstrates that the linear combination of predictors and their regression coefficients is equal to the *logit* or *log odds* (the logarithm of the odds of the outcome). Exponentiation of the *logit* to the base *e* yields the odds of the outcome.

Odds and Odds Ratio

The odds of an event are the probability of its occurrence divided by the probability that it will not occur. The odds ratio is the ratio of one odds to another [8, 43, 99]. In the case of a binary logistic regression model for mortality, the odds ratio for a given predictor variable would be the ratio of the mortality odds in the presence of the predictor divided by the odds of mortality in its absence. This is conveniently obtained by exponentiation of the regression coefficient of the predictor variable to the base *e*. Medical outcome studies typically cite the odds ratios for a number of significant predictor variables along with their 95% confidence intervals.

Overfitting

Overfitting implies an excessively complex model with too many parameters to be estimated from the available data [8]. The resulting model will appear to have an extremely good fit to the training set. However it will generalize poorly to subsequent test samples and have limited ability to predict future events.

Propensity Score

The propensity score is a method to account for selection factors in nonrandomized, observational studies [6, 7]. A logistic regression model is developed using treatment received as the outcome variable. Using this regression equation, the propensity of each patient to receive the particular treatment is calculated. By grouping patients based on their propensity scores (often stratified by quintiles), it is usually possible to obtain treatment and nontreatment groups that are well matched with regard to most predictor variables, although unbalanced with regard to number of subjects. At this point, further comparative analyses between treatment and nontreatment groups can be performed.

Receiver Operating Characteristic (ROC) Curve

The receiver operating characteristic curve was developed in the early days of radar imaging to help separate signals from noise [8, 43, 55]. During the past two decades this technique has been used as a visual method of assessing two naturally competing characteristics of any diagnostic test, sensitivity, and specificity. The receiver operating characteristic curve depicts the specificity and sensitivity of the diagnostic test or predictive model at various cut points (the probability that a researcher requires to assign a patient to the event category). The area under the receiver operating characteristic curve is also called the c-index or c-statistic. It may also be thought of as the percentage of all possible discordant patient pairs (one patient has an event and the other does not) for which the model predicts a higher event probability for the patient who actually has the event. An receiver operating characteristic curve area of 0.5 is no better than a coin toss, whereas an area of 1.0 is perfect discrimination.

Transformations

Transformations are changes in the scale of measurement of a variable during model development. Common reasons for doing this include (1) variance stabilization, (2) linearization, (3) normalization, (4) to simplify handling of the data, and (5) to enable more appropriate presentation of the results [100].

Underfitting

Underfitting is the failure to include one or more important predictor variables, which may lead to poor predictive accuracy.