The Dangers of Gathering Data: Surgeon-Specific Outcomes Revisited

Victor A. Ferraris, M.D., Ph.D.,^{1,2} Suellen P. Ferraris, Ph.D.,¹
Paulette S. Wehner, M.D.,³ and Edward R. Setser, M.D.¹

ABSTRACT

The accuracy of risk adjustment is important in developing surgeon profiles. As surgeon profiles are obtained from observational, nonrandomized data, we hypothesized that selection bias exists in how patients are matched with surgeons and that this bias might influence surgeon profiles. We used the Society of Thoracic Surgeons risk model to calculate observed to expected (O/E) mortality ratios for each of six cardiac surgeons at a single institution. Propensity scores evaluated selection bias that might influence development of risk-adjusted mortality profiles. Six surgeons (four high and two low O/E ratios) performed 2298 coronary artery bypass grafting (CABG) operations over 4 years. Multivariate predictors of operative mortality included preoperative shock, advanced age, and renal dysfunction, but not the surgeon performing CABG. When patients were stratified into quartiles based on the propensity score for operative death, 83% of operative deaths (50 of 60) were in the highest risk quartile. There were significant differences in the number of high-risk patients operated upon by each surgeon. One surgeon had significantly more patients in the highest risk quartile and two surgeons had significantly less patients in the highest risk quartile ($p < 0.05$ by chi-square). Our results show that high-risk patients are preferentially shunted to certain surgeons, and away from others, for unexplained (and unmeasured) reasons. Subtle unmeasured factors undoubtedly influence how cardiac surgery patients are matched with surgeons. Problems may arise when applying national database benchmarks to local situations because of this unmeasured selection bias.

KEYWORDS: Cardiovascular operations, outcomes, complications, risk adjustment

The operative mortality rates for cardiac surgeons are widely disseminated. In some states, they are published in the newspaper—the so-called physician “report cards.”^{1–5} Although these surgeon-specific mortality rates are adjusted for preoperative patient risk factors, it is important that the risk-adjustment models are correct because of the obvious profound effect that publication of surgeon-specific mortality rates may have. The typical form that report cards take is to grade

surgeons by their risk-adjusted operative mortality for one operation, coronary artery bypass grafting (CABG). To grade a provider, the expected number of deaths (E or expected rate) is calculated using some risk-adjustment scheme (usually logistic regression modeling) derived from deaths observed in the entire patient population. The expected mortality rate is then compared with the observed number of deaths for the provider (O or observed rate). This gives the observed

¹Division of Cardiovascular and Thoracic Surgery, Department of Surgery, University of Kentucky Chandler Medical Center, Lexington, Kentucky; ²Department of Surgery, Lexington Veterans Affairs Medical Center, Lexington, Lexington, Kentucky; and ³Division of Cardiology, Marshall University School of Medicine, Huntington, West Virginia.

Address for correspondence and reprint requests: Victor A. Ferraris, M.D., Ph.D., Tyler Gill Professor, Division of Cardiovascular and Thoracic Surgery, A301 Kentucky Clinic, University of Kentucky,

740 South Limestone, Lexington, KY 40536-0284 (e-mail: ferraris@earthlink.net).

Int J Angiol 2011;20:223–228. Copyright © 2011 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel: +1(212) 584-4662.

Published online: July 21, 2011.

DOI: <http://dx.doi.org/10.1055/s-0031-1284433>.

ISSN 1061-1711.

to expected (O/E) ratio or the risk-adjusted mortality rate (RAMR). A confidence limit is assigned to each provider's observed mortality rate, and each provider is compared with the population mean. Providers who are significantly above or below the mean are considered outliers, above the mean being "bad" and below the mean being "good" in simplistic "report card" terms. The biggest single shortcoming of the risk-adjustment methodology is its lack of proven effectiveness in defining quality of care. Even though it may seem obvious that differences in risk-adjusted outcomes reflect differences in quality of care, this is far from proven. Many examples of the shortcomings of risk-adjustment models can be found.^{1,2,6} For example, different risk-adjustment models that presumably deal with similar groups of patients, give different results. The risk-adjustment model of the Society of Thoracic Surgeons (STS) is formulated from data on over 1,000,000 patients.⁷⁻⁹ This model found 27 variables that were predictive of operative mortality. Other similar risk models such as those from Massachusetts¹⁰ and from Toronto¹¹ found far fewer variables that were equally predictive of mortality. If risk-adjustment models of surgeon-specific mortality represent a valid means of comparing quality of care among physicians, then why don't similar models give similar results? Other examples of inadequacies of risk-adjustment methods exist. In fact, Naftel described nine reasons why different investigators with the same dataset could produce different risk models.⁶ Almost all risk-adjustment models are based on observational datasets. This means that risk-adjustment models that rank doctors are based on non-randomized data, and the possibility of selection bias exists. These uncertainties about risk-adjustment methods led us to hypothesize that there is selection bias in how surgeons are matched with patients. As an example, one consequence of this hypothesis might be that high-risk patients are preferentially referred to certain surgeons and that this referral bias alters risk adjustment since referral patterns are not a variable included in the risk-adjustment model. We undertook a study to evaluate the possibility of selection bias in patients undergoing cardiac operations at a single institution.

METHODS

Patient Database

We used a large database that is maintained by one of our hospitals to track outcomes from cardiac surgery. This database contained 2298 patients who had CABG over a 4-year period. Data elements were generated concurrently during the patient's hospitalization, not after discharge. A dedicated database nurse was responsible for gathering the data daily. Examination of the quality of the data revealed that there is both high accuracy and very few elements of missing data.

We calculated the O/E ratio for all surgeons who performed more than 125 operations per year during the study period using the STS risk model. Surgeons were grouped into either high or low mortality groups based on whether their O/E ratio was above or below one.

Statistical Methods

Logistic regression modeling generated independent predictors of operative mortality in the hospital dataset. Forty-three preoperative variables were entered into the regression model in a forward stepwise manner to obtain predictors of operative mortality. Significant predictor variables in the regression model were obtained using "bootstrap bagging" methods to obtain the best estimate of the dependent regression variables.^{12,13}

Statistical methods are available to assess bias in regression models of risk adjustment.¹³⁻¹⁵ One method uses balancing scores to derive comparable groups. We used a particular type of balancing score called the propensity score, to compare outcomes among surgeons in closely matched groups.^{14,15} This methodology identifies selection bias in observational studies and allows more bias-free comparisons between surgeons. We used logistic regression modeling to derive a propensity score for operative death for each patient in the database. Forty-three preoperative patient variables were entered into the logistic model to derive the propensity score. Patients were divided into equal quartiles based on their propensity scores. Within each quartile, we compared outcomes among six different surgeons to compare outcomes among patients with similar risks of mortality.

SPSS statistical software (version 10, IBM/SPSS, Armonk, NY) performed the model building and other statistical calculations. In all cases, a p value of ≤ 0.05 was considered significant.

RESULTS

There were six surgeons who performed more than 125 cardiac operations per year at any time during the study period. The range of O/E ratios for these surgeons is shown in Fig. 1. There were four surgeons in the high-mortality group with an O/E ratio greater than one, and there were two surgeons in the low-mortality group (O/E less than one).

When patients were grouped into quartiles based on the risk of operative mortality (i.e., their propensity score) more than 80% of the operative deaths occurred in Quartile 4 or the highest risk strata (Fig. 2). For comparisons among surgeons this risk strata was the most informative and is referred to as the high-risk quartile.

Multivariate predictors of operative mortality using logistic regression are shown in Table 1. There was

Figure 1 Observed to expected mortality ratios for surgeons who performed more than 125 coronary artery revascularizations, based on the Society of Thoracic Surgeons National Database.

no association between the operating surgeon and operative mortality in this regression model.

There were, however, significant differences among the percentage of high-risk patients operated upon by individual surgeons (Fig. 3). Four surgeons had significantly increased or decreased numbers of high-risk referrals. The surgeon with the highest O/E ratio (Surgeon 1 in Fig. 3) had a significantly greater percentage of high-risk referrals than other surgeons. Two surgeons had significantly fewer high-risk referrals. Interestingly, the surgeon with the lowest (substitute “best”) O/E ratio operated upon significantly fewer high-risk patients (Surgeon 6 in Fig. 3).

We used our database to determine the multivariate predictors of operative mortality. The results of this unmatched, nonrandomized, multivariate comparison indicate that the operating surgeon was not a predictor of operative mortality in this model (Table 1). Conventional risk factors such as preoperative shock, urgency of operation, ejection fraction, and age were more important independent predictors of

Figure 2 Operative deaths in each of 4 propensity matched quartiles.

operative mortality than the operating surgeon (Table 1).

DISCUSSION

These results suggest that high-risk patients are preferentially shunted to certain surgeons, and away from others, for unexplained (and unmeasured) reasons. This is good evidence that selection bias exists within the study group. Unexplained factors result in high-risk patient undergoing operations by specific surgeons for uncertain reasons. Possible reasons include cardiologist referral preference, excessive caution on the part of some surgeons, or multiple other factors. A word of caution is necessary in interpreting these results. The numbers of patients in some of the risk categories are small. The results of this study should be confirmed in a larger database, but these preliminary results suggest that selection bias exists in the study group and in how patients are matched with surgeons.

Table 1 Multivariate Predictors of Operative Mortality by Logistic Regression

Variable	p Value	Odds Ratio	95% CI	
Preoperative shock	<0.000	12.280	4.351	34.658
Ejection fraction (per percent below mean)	<0.000	0.963	0.943	0.983
Preoperative IV nitroglycerin	0.001	2.633	1.457	4.757
Age (per year above mean)	0.003	1.046	1.015	1.077
Preoperative steroid use	0.007	3.678	1.437	9.414
No antiplatelet drug within 24 h of operation	0.011	0.487	0.281	0.846
Female gender	0.027	0.531	0.303	0.929
No preoperative β -blocker	0.031	0.540	0.308	0.945
Peripheral vascular disease	0.037	1.932	1.039	3.593
Coronary artery endarterectomy required	0.040	2.039	1.033	4.025
Constant	0.000	0.014		

Note: Preoperative variables have been entered in the model in a forward stepwise manner, including the surgeon performing the operation as one of the preoperative variables.
IV, intravenous.

Figure 3 Number of patients in the highest risk quartile per surgeon.

It is not surprising that risk adjustment fails to answer definitively the question of “which provider is better” when the providers treat different populations. In reality, some providers do better with certain types of patients and worse with others. In a rational world, providers will concentrate (or be forced to concentrate by referring physicians or other outside forces) on their most successful types of cases. The main protection against being misled is to compare only providers with generally similar patient mixes. When providers are compared, examining patient outcomes separately within each risk stratum helps guard against misinterpretation.¹⁶ Interestingly, the surgeon in our group with the lowest (substitute “best”) O/E ratio was among the least likely to operate upon high-risk patients. In statistical terms, this constitutes important evidence of selection bias within the study group.

One of our mentors said that you should never complain about a problem unless you have a solution. At least three solutions to the problem of comparing surgeons based on O/E ratios come to mind. First, better statistical models for risk adjustment need to be used. The dissemination of surgeon’s mortality rates increased in the last decade. More and more states provide risk-adjusted mortality rates to the lay public. There has not been a similar expansion of the statistical methods used to calculate RAMRs. Statisticians who address this problem believe that there are better ways to do risk stratification.^{17,18} Statistical methodology is available, and has been available for many years, to account for

the incorrect assumptions of standard risk-adjustment models. One of these methods is called hierarchical modeling. This methodology is computer-intensive but provides more accurate comparisons of mortality rates among providers. Other improvements in statistical modeling can and should be applied to the problem.^{17,18} Two incorrect assumptions are used in constructing the O/E ratios usually used to compare providers. The expected mortality rates are assumed to be independent of the observed mortality rates—an incorrect assumption. Furthermore, no sampling error is attached to the expected values—another incorrect assumption. The effect of making these two assumptions is to identify too many outliers (in either direction).¹⁸ Statistical methodology is available, and has been available for many years, to account for these incorrect assumptions. The methodology involves construction of hierarchical regression models. Hierarchy means nesting and the name of these models implies that they incorporate (or nest) other levels of analysis within the analysis of provider mortality. For example, patients are nested within providers, but then patients are also nested within hospital groups (patients treated at a given hospital). Other nestings are possible within the analysis, such as patients referred by a given cardiologist. The most important feature of hierarchical models is that the model recognizes that the nested observations may be correlated (e.g., mortality may depend on the surgeon and the hospital where care is provided and other usually unspecified variables such as referral patterns, academic

status, hospital size and location, etc.). Computer-intensive methods are available to produce hierarchical models for risk-adjusted surgeon mortality. Most statisticians recognize hierarchical models as the preferred method to perform this type of provider analysis but the methods are complex, labor-intensive, and not included in most commercially available computer software.¹ At present, hierarchical regression is the “gold standard” for risk adjustment of dichotomous outcomes and for producing provider report cards. Unfortunately this gold standard is rarely used.

Second, methods of making risk stratification and surgeon profiling less punitive need exploration. Surgeon-specific profiling started from a noble purpose—improving quality of surgical care. Inevitably, the focus on quality took a backseat to rankings among hospitals and surgeons with punitive intent. Efforts should be made to regain the focus on quality of care, rather than punitive profiling. Donabedian and Bashshur suggested that quality in health care is defined as improvement in patient status after accounting for the patient’s severity of illness, presence of comorbidity, and the medical services received.¹⁹ They further proposed that quality could best be measured by considering three domains: structure, process, and outcome. Only recently has the notion of measuring health care quality using this framework been accepted and implemented. In 2000, the Institute of Medicine (IOM) issued a report that was highly critical of the U.S. health care system, suggesting that between 50,000 and 90,000 unnecessary deaths occur yearly because of errors in the health care system.²⁰ The IOM reports created a heightened awareness of more global aspects of quality. For most of the history of cardiac surgery, quality was equated with operative mortality (i.e., outcome measure). After the IOM report appeared, a distinct change in the landscape of quality assessment occurred, and quality evaluation of other aspects of Donabedian and Bashshur’s framework began. The narrow focus on operative mortality gave way to a broader analysis that also included additional outcome measures falling under the general category of operative morbidity. The emphasis on such outcomes measures expanded to include the processes of surgical care delivery. Process measures monitor provider compliance with desirable, often evidence-based, processes and systems of care. These measures typically include the choice of medication, timing of antibiotic administration, use of surgical techniques such as the internal thoracic artery graft, and other interventions considered appropriate for optimal care. Finally, structural measures such as the use of health information technology, physical plant design, participation in systematic clinical registries, and procedural volume are also considered important elements in this more global model of medical quality and performance. All these measures fall under the general rubric of performance measures.

Birkmeyer et al enumerated certain advantages and disadvantages associated with each specific type of performance measure.²¹ For example, the fact that structural measures can be readily tabulated in an inexpensive manner using administrative data are a distinct advantage. On the other hand, many structural measures do not lend themselves to alteration. Particularly in smaller hospitals, there may simply be no way to increase procedural volume or to introduce costly design changes in an attempt to improve their performance on structural measures. Attempts to alter structure might even have adverse consequences (e.g., unnecessary operations, costly and unnecessary new beds). Process measures are often linked to health care quality and they are usually actionable on a practical level. Their major disadvantage lies in the fact that process measures may not be generally applicable to all patients undergoing a given procedure, and their linkage to outcomes may be weak. Outcomes measures are the most important endpoint for patients, but accurate assessment of outcomes is often limited by inadequate sample size and lack of appropriate risk adjustment. Redefining quality based on all three quality measures (structure, process, and outcomes) offers a better chance of improving quality while making surgeon-specific outcome measures less punitive. The STS recognized the shortcomings of using surgeon-specific outcome measures to define quality and instituted a rating scheme that included all three quality measures into a composite score.²² This rating scheme gained national attention and was adopted by the Consumers Union as a way to measure quality in heart surgery programs (<http://www.consumerreports.org/health/doctors-hospitals/surgeon-ratings/heart-surgery-ratings/overview/index.htm>).

Third, the ideal way to improve outcomes is to put it in the realm of quality improvement—similar to what airline pilots do to decrease errors. There are several examples of nonpunitive, peer-based, quality improvement models for cardiac surgery that are used currently. The Northern New England project and the Virginia quality improvement initiative are two examples.^{23,24} Proliferation of these sorts of projects will ultimately result in better acceptance by providers of this important means of assessing outcomes and comparing providers.

To summarize, surgeon-specific profiles contain bias. We suggest that these biased profiles are likely incomplete, possibly wrong, not applicable at the local level, and misleading to the lay public if disseminated. Other means of assessing quality are available and should be used.

NOTE

Parts of this work were presented, in abstract form, at the 42nd Annual AHA Conference on Cardiovascular Disease Epidemiology and Prevention, April 23–26, 2002, Honolulu, HI.

REFERENCES

1. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg* 2001;72(6):2155–2168
2. Grunkemeier GL, Zerr KJ, Jin R. Cardiac surgery report cards: making the grade. *Ann Thorac Surg* 2001;72(6):1845–1848
3. Schneider EC, Epstein AM. Use of public performance reports: a survey of patients undergoing cardiac surgery. *JAMA* 1998;279(20):1638–1642
4. Marshall MN, Shekelle PG, Leatherman S, Brook RH. The public release of performance data: what do we expect to gain? A review of the evidence *JAMA* 2000;283(14):1866–1874
5. Thomas JW, Hofer TP. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Med Care* 1999;37(1):83–92
6. Naftel DC. Do different investigators sometimes produce different multivariable equations from the same data? *J Thorac Cardiovasc Surg* 1994;107(6):1528–1529
7. Edwards FH, Clark RE, Schwartz M. Coronary artery bypass grafting: the Society of Thoracic Surgeons National Database experience. *Ann Thorac Surg* 1994;57(1):12–19
8. Hattler BG, Madia C, Johnson C, et al. Risk stratification using the Society of Thoracic Surgeons Program. *Ann Thorac Surg* 1994;58(5):1348–1352
9. 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(3):903–908
10. Ghali WA, Ash AS, Hall RE, Moskowitz MA. Statewide quality improvement initiatives and mortality after cardiac surgery. *JAMA* 1997;277(5):379–382
11. 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(16):2098–2104
12. Breiman L. Bagging predictors. *Mach Learn* 1996;24:123–140
13. Blackstone EH. Breaking down barriers: helpful breakthrough statistical methods you need to understand better. *J Thorac Cardiovasc Surg* 2001;122(3):430–439
14. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;17(19):2265–2281
15. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Statist Assoc* 1984;79:516–524
16. Iezzoni LI. Risk Adjustment for Measuring Health Care Outcomes. 3rd ed. Chicago: Health Administration Press; 2003
17. Shahian DM, Torchiana DF, Normand SL. Implementation of a cardiac surgery report card: lessons from the Massachusetts experience. *Ann Thorac Surg* 2005;80(3):1146–1150
18. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J R Stat Soc [Series A]* 1996;159:385–443
19. Donabedian A, Bashshur R. An Introduction to Quality Assurance in Health Care. Oxford; New York: Oxford University Press; 2003
20. Kohn LT, Corrigan J, Donaldson MS. Institute of Medicine. (U.S.). Committee on Quality of Health Care in America. To Err is Human: Building a Safer Health System. Washington, D.C.: National Academy Press; 2000
21. Birkmeyer JD, Dimick JB, Birkmeyer NJ. Measuring the quality of surgical care: structure, process, or outcomes? *J Am Coll Surg* 2004;198(4):626–632
22. Shahian DM, Edwards FH, Ferraris VA, et al. Society of Thoracic Surgeons Quality Measurement Task Force. Quality measurement in adult cardiac surgery: part 1—conceptual framework and measure selection. *Ann Thorac Surg* 2007;83(4, Suppl):S3–S12
23. O'Connor GT, Plume SK, Olmstead EM, et al. The Northern New England Cardiovascular Disease Study Group. A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. *JAMA* 1996;275(11):841–846
24. Grover FL, Johnson RR, Shroyer AL, Marshall G, Hammermeister KE. The veterans affairs continuous improvement in cardiac surgery study. *Ann Thorac Surg* 1994;58(6):1845–1851