Off the Shelf or Recalibrate? Customizing a Risk Index for Assessing Mortality

(#2003-51630 . . . April 18, 2003)

James F. Reed III, PhD,[1] Stephen A. Olenchock Jr, DO,[2] Sabina A. Murphy, Fernando M. Garzia, MD[3]

[1]Research Institute, [2]Department of Surgery, and [3]Department of Surgery, Division of Cardiovascular Surgery, St. Luke's Hospital & Health Network, Bethlehem, Pennsylvania, USA

Dr. Reed

ABSTRACT

Background: Public "report cards" for cardiac surgery have been freely available from a variety of sources. These risk-adjusted indices serve as a means of benchmarking outcomes performances, allowing comparisons of outcomes between surgical programs, and quantifying quality improvement programs. We examined two alternative strategies for using previously developed risk-adjusted mortality models in a community hospital: (1) using the model "off the shelf" (OTS) and (2) recalibrating the existing model (RM) to fit the institution-specific population.

Methods: Six OTS models were used: Parsonnet (PA), Canadian (CA), Cleveland (CL), Northern New England (NNE), New York (NY), and New Jersey (NJ). The RM models were created by each model's independent variables and definitions and adjusting the weighting with logistic regression methods. The accuracy, the C statistic, and the precision of each model were assessed for in-hospital mortality. We compared the OTS version of each model to the RM version with methods detailed by Hanley and McNeil.

Results: The RM C statistic was improved for all risk-adjusted models, most notably in the statistical improvement seen in the PA (0.053 improvement) and NJ (0.052 improvement) indices. Statistical gains in precision were also seen in the RM models for the PA, CL, and NNE indices. Conversely, one model, the CA model, was more poorly calibrated in the RM model compared with the OTS model, despite an improved C statistic (0.062).

Conclusions: The RM strategy provides institution-explicit models that demonstrate a higher degree of accuracy and precision than the OTS models.

INTRODUCTION

Public "report cards" or "provider profiles" for cardiac surgery, both institution specific and surgeon specific, have been freely available from a variety of sources, either directly from independent organizations or from their Web sites. These reports are based on various scoring systems or risk-adjusted algorithms for mortality after coronary artery bypass (CAB) graft surgery. In general, these algorithms identify and weight patient characteristics that affect the probability of specific outcomes [Parsonnet 1989, Higgins 1992, O'Conner 1992, Hannan 1994, Tu 1995, Roques 1995, NJDHSS 1998].

Most would agree that comparing outcomes is meaningful only when those outcomes are adjusted for potential differences in the type or case mix of patients. Risk adjustment methods that "level the playing field" use statistical modeling that adjusts for individual patient risk factors in predicting the outcome event of interest. These risk-adjusted prediction models allow for the computation of a provider's expected outcome event rate compared with observed outcome events [Pons 1997]. These risk-adjusted indices also serve as a means of benchmarking outcomes performances, allowing comparison of outcomes between surgical programs, and quantifying quality improvement programs by comparing outcomes from year to year or after a change in surgical practice [Pons 1997].

This study examined two alternative strategies for using already developed risk-adjusted mortality models in a community hospital: (1) using the model "off the shelf" (OTS) and (2) recalibrating the existing model (RM) to fit the institution-specific population.

METHODS

This study was conducted at a community hospital in eastern Pennsylvania. Included were 1135 consecutive patients undergoing CAB surgery from January 1, 1998, to December 31, 1999. Clinical and physiological data were collected by a single trained cardiac data abstractor following an abstraction protocol and the data definitions of the Society for Thoracic Surgeons. Data were entered into a computerized database, and risk scores and predicted hospital mortality rates were calculated with the models recommended in published reports for each of the risk-adjusted models. The outcome measure was in-hospital mortality.

Six risk adjustment models were selected that estimated mortality risk following CAB surgery: the Parsonnet (PA) [Parsonnet 1989], Canadian (CA) [Tu 1995], Cleveland (CL) [Higgins 1992], New York (NY) [Hannan 1994], Northern

Received April 3, 2003; accepted April 18, 2003.

Address correspondence and reprint requests to: James F. Reed III, PhD, Research Institute, St. Luke's Hospital & Health Network, 801 Ostrum St, Bethlehem, PA 18015, USA; 1-610-954-4942; fax: 1-610-954-4979 (e-mail: ReedJ@slhn.org).

New England Cardiovascular (NNE) [O'Conner 1992], and New Jersey (NJ) [NJDHSS 1998] models.

The choice of these models was guided by theoretical and practical reasons. In general, predictive models perform better in their original setting than when transported to other patient populations. Some of that unpredictability may be due to differences in outcome measurement or data collection. To minimize this possibility, we used the definitions of each variable in each model and matched those definitions to the variable definitions in our cardiac surgery database. Another potential bias is that the original predictive models may be overfitted to the study population and therefore may become less able to be generalized to other populations. This issue is addressed with traditional logistic regression analyses by restricting the number of independent variables to avoid the potential for overfitting.

The PA mortality model was originally developed with 3500 consecutive patients who underwent CAB and/or valve surgery between 1982 and 1987 [Parsonnet 1989]. The outcome variable in the PA model was mortality occurring within 30 days of surgery. The resulting additive model contained 11 variables. The PA model has been examined worldwide in a variety of settings with sample sizes ranging from 465 to 3443 (Weightman 1997, Pliam 1997, Iezzoni 1998, Martinez-Alario 1999, Geissler 2000]. The present study used a current version of the PA mortality model, the St. George's Hospital PA calculator [Martinez-Alario 1999] (Table 1).

The CA score was developed from the Ontario Ministry of Health Provincial Adult Cardiac Care Network computerized registry [Tu 1995]. Clinical information from 9 institutions in Ontario was gathered and entered into this database at the time of referral for cardiac surgery. The CA score was derived with data from 6213 patients and validated for 6885 patients who had undergone cardiac surgery. The outcomes of interest were in-hospital mortality, a very long stay in the intensive care unit (≥6 days), and a very long postoperative stay (≥17 days). Three logistic regression models were then developed, one for each of the outcome variables. The CA risk index was then created by rounding the mean of the 3 odds ratios for each risk factor in the different logistic models to the nearest integer. The resulting risk index for mortality included 6 factors (Table 1). The CA model has also been tested in a variety of settings with sample sizes ranging from 505 to 7491 [Weightman 1997, Iezzoni 1998, Lawrence 2000]. For this study of isolated CAB outcomes, 1 variable (type of surgery) was set to its null value, leaving only 5 variables and their associated risk scores.

The CL severity score relates mortality (death during a hospitalization for surgery, regardless of length of stay or within 30 days of hospital discharge) and morbidity (cardiac complication, prolonged ventilation, central nervous system complication, oliguric or anuric renal failure, serious infection, and death) to preoperative severity of illness in patients undergoing CAB [Higgins 1992]. The CL mortality index identified 13 preoperative risk factors, assigned a weight to each risk factor, and related a total score to postoperative mortality. For this study, the 9-variable algorithm of Peterson and colleagues was used [Peterson 2000] (Table 1).

The NY risk-adjusted model was developed with patient data from 30 hospitals that performed 57,187 isolated CAB surgeries from 1989 through 1992 [Hannan 1994]. The NY risk-adjusted model served as a benchmark marker in these hospitals' quality improvement initiative to assess changes in outcomes of CAB surgery in New York via the collection, analysis, and dissemination of information regarding risk factors, mortality, and complications of CAB surgery. The NY index has been used extensively and has performed well as an external benchmark. It is an algorithm developed in one population and applied to the evaluation population in subsequent analyses [Orr 1995, Ivanov 1999, Peterson 2000]. The 14-variable mortality model and multivariate odds ratios were used for this study, as shown in Table 1 [Orr 1995].

The NNE model was developed with data for 3055 patients from 5 clinical centers [O'Conner 1992]. These centers' data set was randomly divided into a derivation set that was used to develop the prediction model and a validation set that was used to assess the performance of the model. The dependent variable was in-hospital mortality. Logistic regression analysis was used to derive multivariate odds ratios to predict the probability of an in-hospital death conditional on patient preoperative characteristics. This study uses the 1999 version of the NNE mortality model [DeLong 1997] (Table 1).

Finally, the present study used the NJ model developed by the Department of Health and Senior Services in 1998. This model uses data from 8377 patients who underwent isolated CAB surgery in 14 hospitals located throughout New Jersey [NJDHSS 1998]. All hospitals in New Jersey collect detailed patient-specific information related to the patient's demographics, preoperative risk factors, and complications of surgery and discharge status (died in the hospital, discharged alive) with a single data collection system (the Society of Thoracic Surgeons system) (Table 1).

A major requisite of any predictive model is that it adequately describe the risk in the study population. Each model was assessed in terms of 2 standard measures of a model's performance: discrimination and calibration. Discrimination is the ability of a model to correctly distinguish between patients who will die from those who will survive. Model discrimination was assessed with the C statistic or the area under the receiver operating curve [Hanley 1982], and comparisons of the OTS and RC models were completed with methods described by Hanley and McNeil [Hanley 1983].

Calibration is a measure of the model's ability to predict survival for various levels of patient risk and is measured by the Hosmer-Lemeshow statistic (H-L χ^2). The H-L χ^2 statistic is a variation of the χ^2 statistic that compares the observed and predicted outcomes for 10 equally populated levels of patient risk. A poorly calibrated model has an H-L χ^2 *P* value of <.05 [Lemeshow 1982].

Our remodeling algorithm consisted of performing a logistic regression analysis with the variables identified by each OTS model. The weight for each of the variables was then assigned the new adjusted odds ratios. A receiver operating curve analysis was then completed with the RM model.

© 2003 Forum Multimedia Publishing, LLC

Table 1. Coronary Artery Bypass Mortality Risk Indices*

Parsonnet Risk Index	Risk Score
Female sex	1
Morbid obesity	3
Diabetes (type unspecified)	3
Hypertension (systolic blood pressure >140 mm Hg)	3
Left ventricular ejection fraction	
Good (>49%)	0
Fair (30%-49%)	2
Poor (<30%)	4
Age	
70-74 y	7
75-79 y	12
>79 y	20
Reoperation	
First	5
Second	10
Preoperative intra-aortic balloon pump	2
Left ventricular aneurysm	5
Emergency surgery following coronary angioplasty	10
Dialysis dependency	10

Canadian Risk Index	Risk Score
Age	
<65 y	0
65-74 y	2
≥75 y	3
Female sex	1
Left ventricular function	
Grade 1 (>50%)	0
Grade 2 (35%-50%)	1
Grade 3 (20%-34%)	2
Grade 4 (<20%)	3
Type of surgery	
CAB only	0
Single valve	2
Multivalve or CAB plus valve	3
Urgency of surgery	
Elective	0
Urgent	1
Emergency	4
Repeat operation (previous CAB)	2

Cleveland Risk Index	Risk Score
Age	
65-74 y	1
≥75 y	2
Cerebrovascular disease	1
Chronic pulmonary disease	2
Anemia (hematocrit ≤0.34)	2
Renal insufficiency	
Serum creatinine 1.6-1.8 mg/dL	1
Serum creatinine ≤1.9 mg/dL	4
Severe left ventricular dysfunction (<35%)	3
Operative mitral valve insufficiency	3
Prior CAB surgery	3
Emergency CAB surgery	6

Table 1. Continued

New York	Odds Ratio
Age	
Age, y	1.04
Age ≥70 y	1.05
Female sex	1.52
Coronary disease, left main stenosis >90%	1.43
Reversible ischemia, unstable angina	1.42
Left ventricular function	
<20%	4.06
20%-29%	2.21
30%-39%	1.63
Missing	1.62
Previous myocardial infarction within 7 days	1.69
Preoperative intra-aortic balloon pump	1.39
Congestive heart failure	1.77
Disasters	3.98
Diabetes	1.50
Morbid obesity	1.49
Chronic pulmonary disease	1.36
Dialysis dependence	2.80
Other, previous open heart operation	3.73

Northern New England	Score
Age	
60-69 y	2
70-79 y	3
≥80 y	5
Female sex	1.5
Ejection fraction <40%	1.5
Urgent surgery	2
Emergency surgery	5
Prior CAB	5
Peripheral vascular disease	2
Dialysis or creatinine ≥2 mg/dL	4
Chronic obstructive pulmonary disease	1.5

New Jersey	Odds Ratio
Age	
65-69 y	1.791
70-74 y	2.080
75-79 y	2.964
80-84 y	3.594
≥85 y	4.285
Female sex	1.940
Congestive heart failure	1.611
Renal failure with dialysis	7.600
Renal failure without dialysis	4.009
Lung disease	1.699
Left ventricular function	
1%-20%	2.581
30%-49%	1.555
Cardiogenic shock	2.281
Prior CAB	2.884
Triple-vessel disease	1.428

*CAB indicates coronary artery bypass grafting.

Table 2. Summary Statistics Comparing the Off-the-Shelf Model and the Recalibrated Model*

	OTS Model			Recalibrated Model			
Model	C ± SE	H-L χ^2	*P*	C ± SE	H-L χ^2	*P*	HM *P*
PA	0.752 ± 0.040	4.95	.763	0.805 ± 0.040	4.31	.828	.002
CA	0.693 ± 0.044	1.62	.899	0.755 ± 0.039	13.8	.054	.098
CL	0.748 ± 0.036	12.0	.035	0.769 ± 0.036	5.19	.520	.135
NY	0.735 ± 0.046	10.2	.249	0.768 ± 0.042	11.9	.156	.108
NNE	0.772 ± 0.035	12.1	.145	0.803 ± 0.031	8.44	.391	.067
NJ	0.787 ± 0.031	6.08	.638	0.839 ± 0.026	8.58	.379	.005

*OTS indicates off the shelf; C, C statistic; H-L χ^2, Hosmer-Lemeshow statistic; HM, Hanley-McNeil comparison of OTS and recalibrated models; PA, Parsonnet; CA, Canadian; CL, Cleveland; NY, New York; NNE, Northern New England; NJ, New Jersey.

RESULTS

There were 1135 CAB operations performed at this institution in the 2-year period between January 1, 1998, and December 31, 1999. There were 32 deaths, yielding an overall mortality rate of 2.8%. There were 346 female patients (30.5%) and 789 male patients (69.5%), and the mean age was 67.1 ± 10.3 years.

The Hanley-McNeil comparisons of the OTS and RM models are summarized in Table 2. The RM C statistic was improved for all risk-adjusted models, most notably in the statistical improvement seen in the PA (0.053 improvement) and NJ (0.052 improvement) indices. Statistical gains in precision were also seen in the RM models for the PA, CL, and NNE indices. Conversely, one model, the CA model, was more poorly calibrated in the RM model compared with the OTS model despite an improved C statistic (0.062).

COMMENT

Comparing hospital-specific or surgeon-specific mortality/morbidity rates remains a challenge. Public report cards have used a variety of risk adjustment algorithms or indices for mortality, morbidity, and length of hospital stay following CAB surgery. These indices are typically used retrospectively to adjust for case mix differences between surgeons and institutions in compiling performance profiles. Because patient populations differ significantly between institutions and countries, it is intuitively evident that a comparison of absolute mortality rates is not justified [Parsonnet 1989, Higgins 1992, Peterson 2000]. These risk-adjusted indices were developed to correct for case mix differences in a patient population and allow for comparisons within an institution.

We compared two strategies for assessing the utility of risk-adjusted in-hospital mortality in CAB surgery: (1) the off-the-shelf use of an established risk index and (2) recalibrating that index to ensure a better fit to an institution. If the goal of the use of a risk-adjusted index is to determine trends in mortality over time, then an OTS risk model can be used. A limitation of this strategy is that the reporting of risk factors ("upcoding") may lead to spurious decisions that the risk-adjusted outcomes are improving. Critics have contended that this phenomenon, rather than public report cards, explains the observations of improved outcomes of CAB surgery in New York State [Green 1995]. If the goal of an outcome analysis is contemporary quality improvement, our analysis suggests that institutions should consider recalibrating an existing index to institution-specific patient characteristics. Recalibration generally allows an institution to focus their data collection efforts on a limited number of specific variables. Through recalibration of these variables and a fine-tuning of the risk-adjusted index, the RM index may detect subtle shifts in case mix and outcomes over time or as the index is applied to institutions other than those from which it was derived. For those clinicians or institutional managers who have been using an index that has performed well in their population, periodic recalibration of that index may be sufficient to support institution-specific quality improvement initiatives.

The added benefit to a community hospital in using a RM risk adjustment model is the ability of that institution to establish localized benchmarks for comparing differences between the observed and predicted outcomes as a means for improving performance.

REFERENCES

DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. 1997. Comparing risk-adjustment methods for provider profiling. Stat Med 16:2645-64.

Geissler HJ, Holzl P, Marohl S, et al. 2000. Risk stratification in heart surgery: comparison of six score systems. Eur J Cardiothorac Surg 17:400-6.

Green J, Whitfeld N. 1995. Report cards on cardiac surgeons: assessing New York State's approach. N Engl J Med 332:1229-32.

Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36.

Hanley JA, McNeil BJ. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839-43.

Hannan EL, Kilburn H Jr, Racz M, Shields E, Chassin MR. 1994. Improving the outcomes of coronary artery bypass surgery in New York State. JAMA 271:761-6.

Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Paranandi L. 1992. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients: a clinical severity score. JAMA 267:2344-8.

© 2003 Forum Multimedia Publishing, LLC

Iezzoni LI, Ash AS, Shwartz M, Landon BE, Mackiernan YD. 1998. Predicting in-hospital death from coronary artery bypass graft surgery: do different severity measures give different predictions? Med Care 36:28-39.

Ivanov J, Tu JV, Naylor CD. 1999. Ready-made, recalibrated, or remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation 99:2098-104.

Lawrence DR, Valencia O, Smith EEJ, Murday A, Treasure T. 2000. Parsonnet score is a good predictor of the duration of intensive care unit stay following cardiac surgery. Heart 83:429-32.

Lemeshow S, Hosmer DW Jr. 1982. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 115:92-106.

Martinez-Alario J, Tuesta ID, Plasencia E, Santana M, Mora ML. 1999. Mortality prediction in cardiac surgery patients: comparative performance of Parsonnet and general severity systems. Circulation 99:2378-82.

[NJDHSS] New Jersey Department of Health and Senior Services, Division of Health Care Systems Analysis. Coronary artery bypass graft surgery in New Jersey 1998. Available at: http://www.state.nj.us/health/hcsa/cabgs99/technical.htm. Accessed April 30, 2003.

O'Conner GT, Plume SK, Olmstead EM, et al. 1992. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery: Northern New England Cardiovascular Disease Study Group. Circulation 85:2110-8.

Orr RK, Maini BS, Sottile FD, Dumas EM, O'Mara P. 1995. A comparison of four severity-adjusted models to predict mortality after coronary artery bypass graft surgery. Arch Surg 130:301-6.

Parsonnet V, Dean D, Berstein AD. 1989. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation 79:I3-12.

Peterson ED, DeLong ER, Muhlbaier LH, et al. 2000. Challenges in comparing risk-adjusted bypass surgery mortality results: results from the Cardiovascular Project. J Am Coll Cardiol 36:2174-84.

Pliam MB, Shaw RE, Zapolanski A. 1997. Comparative analysis of coronary surgery risk stratification models. J Invasive Cardiol 9:203-22.

Pons JM, Granados A, Espinas JA, Borras JM, Martin I, Moreno V. 1997. Assessing open heart surgery mortality in Catalonia (Spain) through a predictive risk model. Eur J Cardiothorac Surg 11:415-23.

Roques F, Nashef SA, Michel P, et al. 1999. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg 15:816-23.

Tu JV, Jaglal SB, Naylor CD. 1995. Multicenter validation of a risk index for mortality, intensive care unit stay, and overall hospital length of stay after cardiac surgery: Steering Committee of the Provincial Adult Cardiac Care Network of Ontario. Circulation 91:677-84.

Weightman WM, Gibbs NM, Sheminant MR, Thackray NM, Newman MA. 1997. Risk prediction in coronary artery surgery: a comparison of four risk scores. Med J Aust 166:408-11.