Reliability of risk algorithms in predicting early and late operative outcomes in high-risk patients undergoing aortic valve replacement

Todd M. Dewey, MD^a, David Brown, MD^a, William H. Ryan, MD^a, Morley A. Herbert, PhD^b, Syma L. Prince, RN^a, and Michael J. Mack, MD^a

Objective: Risk algorithms were used to identify a high-risk population for transcatheter aortic valve implantation instead of standard aortic valve replacement in patients with aortic stenosis. We evaluated the efficacy of these methods for predicting outcomes in high-risk patients undergoing aortic valve replacement.

Methods: Data were collected on 638 patients identified as having isolated aortic valve replacement between January 1, 1998 and December 31, 2006, using The Society of Thoracic Surgeons (STS) database. Long-term survival was determined from the Social Security Death Index or family contact. Operative risk was calculated using the STS Predicted Risk of Mortality, the EuroSCORE logistic and additive algorithms, and the Ambler Risk Score. Patients at or above the 90th percentile of risk (8.38% for STS, 33.47% for logistic, 12% for additive, 14.3% for Ambler) were identified as high risk. We then compared actual with predicted mortality and each algorithm's ability to identify patients with the worst long-term survival.

Results: Operative mortality was 24 of 638 (3.76%). An additional 121 (19.0%) patients died during the follow-up study period (mean 4.2 ± 2.7 years). Overall mortality was 145 of 638 (22.7%). Expected versus observed mortality for the high-risk group by algorithm was 13.3% versus 18.8% for STS, 50.9% versus 15.6% for logistic, 14.0% versus 11.9% for additive, and 19.0% versus 13.4% by Ambler. Long-term mortality, per high-risk group, was 64.1% in the STS Predicted Risk of Mortality, 45.3% in the logistic, 45.2% in the additive, and 40.2% in Ambler Risk Score. Logistic regression showed that the STS algorithm was the most sensitive in defining the patients most at risk for long-term mortality.

Conclusion: The STS Predicted Risk of Mortality most accurately predicted perioperative and long-term mortality for the highest risk patients having aortic valve replacement.

From the Cardiopulmonary Research Science and Technology Institute (CRSTI)^a and the Medical City Dallas Hospital^b, Dallas, Tex.

Read at the Eighty-seventh Annual Meeting of The American Association for Thoracic Surgery, Washington, DC, May 5-9, 2007.

Received for publication May 4, 2007; revisions received Aug 29, 2007; accepted for publication Sept 12, 2007.

Address for reprints: Todd M. Dewey, MD, 7777 Forest Ln, Suite A323, Dallas, TX 75230 (E-mail: tdewey@CSANT.com).

J Thorac Cardiovasc Surg 2008;135:180-7
0022-5223/\$34.00

Copyright © 2008 by The American Association for Thoracic Surgery

doi:10.1016/j.jtcvs.2007.09.011

Aortic stenosis (AS) is the most common valvular abnormality encountered in the United States, with a reported incidence of approximately 5 of every 10,000 adults.¹ The prevalence of AS is expected to markedly increase as the US population ages, with aortic valve calcification and stenosis affecting the health of larger numbers of patients.²⁻⁴

Abbreviations and Acronyms

AES	= EuroSCORE additive
ARS	= Ambler Risk Score
AS	= aortic stenosis
AVR	= aortic valve replacement
CI	= confidence interval
LES	= EuroSCORE logistic
STS	= The Society of Thoracic Surgeons
STS-PROM	= STS Predicted Risk of Mortality

Aortic valve replacement (AVR) has been the gold standard for the treatment of critical AS for decades, and the indications for its implementation have been well defined in the American College of Cardiology/American Heart Association 2006 Practice Guidelines for the Management of Patients with Valvular Heart Disease.⁵

Despite unequivocal data showing the efficacy of AVR for critical AS, there are indications that a significant number of patients may not be referred for surgical therapy. A recent report from Varadarajan and colleagues⁶ from Loma Linda reported that only 287 (38%) patients of 740 with AS, defined as a valve area of less than 0.8 cm² by echocardiography, were referred for AVR. Additionally, the Euro Heart Survey on Valvular Heart Disease, a prospective survey of 5001 patients conducted at 92 centers across Europe, reported that 31.8% of patients did not undergo surgical intervention.⁷ Reasons given for nonoperative management in both studies included “old age,” severe comorbidities, and patient refusal.

Transcatheter alternatives to standard AVR, with either self-expanding or balloon expandable valves, have been developed to reduce the anticipated mortality and morbidity of therapy in patients in whom the risk of conventional surgery for critical AS was considered to be too high. Early published data on procedural mortality and morbidity have demonstrated acceptable 30-day results.⁸ Patients with anticipated diminished operative survival, and by extension likely reduced long-term survival, comprise a logical treatment group in which to evaluate technology whose efficacy and in vivo durability are undetermined but offers the chance for potential decreases in morbidity and mortality. Fundamental to this process, however, is the ability to identify and select patients with an anticipated operative mortality with conventional surgery that would be considered prohibitive. Operative risk scoring algorithms are currently being used to identify the appropriate patient population for transcatheter therapy. However, the accuracy of these risk models in identifying high-risk patients appropriate for nonstandard valve therapy has not been fully validated. We evaluated four common risk scoring algorithms for their reliability in identifying high-risk patients undergoing isolated AVR for critical AS based on their accuracy

in predicting operative outcomes and the ability to identify patients most likely to die during longer term follow-up.

Patients and Methods

Between January 1998 and December 2006, 638 patients were identified as having undergone isolated AVR for critical AS. The study received exempt status from the North Texas Institutional Review Board at Medical City Dallas Hospital. The preoperative and perioperative data were prospectively collected and retrospectively extracted from a Society of Thoracic Surgeons (STS) certified database maintained by the Cardiopulmonary Research Science and Technology Institute, a 501C(3) medical research organization that collects clinical data for a 22-surgeon practice covering 18 hospitals in the North Texas region.

The STS Predicted Risk of Mortality (STS-PROM) for aortic valve procedures was calculated when patient data were entered into the STS database, whereas the EuroSCORE algorithms (logistic [LES] and additive [AES]) were calculated according to published guidelines (<http://www.euroscore.org/>). The STS-PROM uses 24 variables of more than 50 total risk factors collected by the algorithm to predict mortality for valve procedures. The LES and AES collect the same 17 variables divided into three categories (patient-related factors, heart-related factors, and operation-related factors) but use differing methods of weighting to arrive at predicted risk. The AES uses an integer approach to calculate risk, whereas the LES uses a logarithmic approach. The Ambler score was also calculated for each patient according to previously published guidelines.⁹ The Ambler Risk Score (ARS) uses an integer approach with a conversion table to translate the additive score into a predicted risk. Patients' risk factors are assigned points on the basis of 13 broad categories of preoperative variable ranging from age to the existence of diabetes and hypertension.⁹

The 90th percentile was calculated for each scoring system and then patients were allocated to a “high-risk” group if their score was at or above the 90th percentile. Inclusion into the high-risk cohort was variable depending on the risk model used to calculate predicted operative mortality.

Perioperative deaths (mortality within 30 days of the operation or during the same admission) were actively tracked from the STS database and late mortality was obtained by searching the Social Security Death Index or through direct telephone contact.¹⁰ Long-term follow-up for mortality was available for all patients. All statistical analysis was carried out with SAS 9.1.3 software (SAS Institute, Inc, Cary, NC).

Results

A total of 638 patients were identified as having undergone isolated AVR for critical AS between January of 1998 and December of 2006. Predicted mortality for the entire group was determined by each risk model and ranged from 4.26% for the STS-PROM algorithm, to 7.46% for the AES, 6.99% for the ARS, and 13.21% for the LES. Operative mortality for the group as a whole was 24 (3.76%) of 638 patients. Patients were followed up for longer-term mortality by the Social Security Death Index or direct family contact. The average length of long-term follow-up was 4.2 ± 2.7 years

TABLE 1. Risk variables used by algorithms

	STS	LES and AES	ARS
Age	X	X	X
Sex	X	X	X
Height	X		
Weight	X		
Body mass index			X
Diabetes	X		X
Chronic lung/pulmonary disease	X	X	
Extracardiac arteriopathy		X	
Peripheral vascular disease	X		
Neurologic dysfunction disease		X	
Previous cardiac surgery		X	
Number of previous operations	X		X
Previous CABG	X		
Previous valve surgery	X		
Renal failure	X		X
Dialysis-dependent renal failure	X		X
Serum creatinine	X	X	X
Hypertension	X		X
Active endocarditis		X	
Type	X		
Cerebrovascular accident	X		
Arrhythmias			X
Ejection fraction			X
Critical preoperative state		X	
Unstable angina		X	
Left ventricular dysfunction		X	
Recent myocardial infarct		X	
Timing	X		
Cardiogenic shock	X		
NYHA classification	X		
Inotropic agents	X		
Pulmonary hypertension (>60 mm Hg)		X	
Pulmonary artery mean pressure	X		
Procedure status	X		X
Emergency/emergency surgery		X	
Intra-aortic balloon pump timing	X		
Other than isolated CABG		X	
Surgery on thoracic aorta		X	
Mitral valve surgery	X		X
Aortic and mitral			X
Concomitant CABG surgery			X
Concomitant tricuspid surgery			X
Postinfarct septal rupture		X	

STS, The Society of Thoracic Surgeons; LES, EuroSCORE logistic; AES, EuroSCORE additive; ARS, Ambler Risk Score; CABG, coronary artery bypass grafting; NYHA, New York Heart Association.

(median 4.11 years). During the follow-up period an additional 121 (19.0%) patients died, for an overall mortality of 145 (22.7%) of 638. **Table 1** summarizes the variables included in each risk model and highlights the differences between each algorithm

The study population was subdivided on the basis of calculated operative risk by the STS-PROM, the AES and

LES, and the ARS. **Table 2** shows the demographic and preoperative variables of the high-risk patients as identified by the different risk models. A high-risk cohort was defined as those patients in the top 10% of predicted risk by each of the four risk models (**Table 3**). All other patients were designated as low risk. The break point for the top 10% of risk was identified by the STS-PROM as those patients having an operative risk greater than 8.38% and by the LES, AES, and ARS as greater than 33.47%, 12.0%, and 14.3%, respectively. Differences in the variables collected between models, and in weighting between the LES and AES algorithms, resulted in different numbers of patients being identified as in the top 10% of risk per model. Additionally, scoring systems derived from integer scales (AES and ARS) resulted in many patients with tied scores, leading to larger numbers in the group at or above the 90th percentile as calculated by score. The STS-PROM and LES both identified 64 patients as being in the top 10% of risk, whereas the AES and ARS identified 83 and 97 patients, respectively. Although many of the same patients identified as high risk were included in each of the cohorts, the differences in the percentage of common patients further illustrates the uniqueness of each model (**Figure 1**). For example, when the STS-PROM cohort was compared with the LES, only 62.5% of the patients were in both groups. Likewise, only 70.3% and 60.9% of the STS-PROM patients were also found in the AES and ARS, respectively, and only 4.6% (29 patients) were in the top 10% of all three scoring systems—STS, LES, and ARS, the concordant group. By the same risk variables, all LES patients were included in the AES group, but the absolute numbers identified as high risk by each model differed owing to the integer method used by the AES.

Average age was 79.5 ± 8.9 years in the STS-PROM group and 79.3 ± 6.3 , 78.6 ± 7.4 , and 77.3 ± 7.7 years in the LES, AES, and ARS groups, respectively (**Table 2**). The average ejection fraction ranged from a low of $41.1\% \pm 14.8\%$ in the ARS cohort up to a high of $45.7\% \pm 14.6\%$ in the STS-PROM. Significant differences were seen between the groups in terms of the distribution of patients with a history of cerebrovascular accident, peripheral vascular disease, and cerebrovascular disease. The risk variable of renal failure trended toward significance, with 26.6% of the STS-PROM having this risk factor versus 20.3%, 19.1%, and 10.3% in the LES, AES, and ARS groups, respectively ($P = .062$). There were no differences among the groups in terms of percentage of patients having emergency procedures, redo sternotomies, chronic pulmonary disease, New York Heart Association class III-IV heart failure, or having had a previous myocardial infarction. Likewise, the distribution of women in each group was not significantly different, ranging from 54.7% in the STS-PROM to 45.3%, 47%, and 46.4%, respectively, in the LES, AES, and ARS cohorts.

TABLE 2. Preoperative demographics of high-risk patients identified by various risk algorithms

	STS-PROM (64) %	LES (64) %	AES (84) %	ARS (97) %	P value
Age (y)	79.5 + 8.9	79.3 + 6.3	78.6 + 7.4	77.3 + 7.7	.22
Ejection fraction	45.7 + 14.6	43.0 + 14.1	43.7 + 14.2	41.1 + 14.8	.31
Female gender	54.7	45.3	47.6	46.4	.70
Diabetes	42.2	26.6	31.0	41.2	.13
Renal failure	26.6	20.3	19.1	10.3	.062
Renal failure requiring dialysis	13.2	8.8	9.8	7.3	.85
Cerebrovascular accident	22.2	38.1	28.9	13.5	.004
Chronic lung disease	23.1	12.5	15.8	7.7	.82
Peripheral vascular disease	34.4	37.5	35.7	15.5	.004
Cerebrovascular disease	29.7	40.6	38.1	18.6	.008
Previous CABG	44.4	57.1	51.8	41.7	.22
Myocardial infarction	18.8	25.0	23.8	16.5	.49
Congestive heart failure	75.0	67.2	64.3	65.0	.51
Cardiogenic shock	6.3	4.7	3.6	3.1	.78
Arrhythmia	35.9	32.8	32.1	42.3	.48
NYHA class (III-IV)	71.9	60.9	63.1	64.6	.33
Status					
Elective	31.3	53.1	56.0	43.3	.075
Emergency	1.6	1.6	1.2	3.1	
Urgent	67.2	45.3	42.9	53.6	
Intra-aortic balloon pump	14.1	17.2	16.7	11.5	.70

STS-PROM, The Society of Thoracic Surgeons Predicted Risk of Mortality; LES, EuroSCORE logistic; AES, EuroSCORE additive; ARS, Ambler Risk Score; CABG, coronary artery bypass grafting; NYHA, New York Heart Association.

Mean predicted operative mortality for the STS-PROM high-risk group was 13.31% compared with 50.87%, 14.04%, and 19.03% for the LES, AES, and ARS high-risk cohorts, respectively (Table 3).

Measured operative mortality was 18.75% (12/64) in the STS-PROM high-risk cohort, 15.63% (10/64) and 11.90% (10/84) in the LES and AES groups, respectively, and 13.40% (13/97) in the ARS group. The observed/expected ratios calculated for perioperative mortality were 0.31, 0.85,

and 0.71 in the LES, AES, and ARS cohorts, respectively. The STS-PROM group had an observed/expected mortality ratio of 1.41. Longer-term follow-up showed that 45.3% (29/64) of the patients identified by the STS-PROM as being at high risk died within the follow-up period. Within the LES and AES groups, 29.7% (19/64) and 33.3% (28/84), respectively, of the patients identified as being at high risk died during the follow-up period. Only 26.8% (26/97) of the ARS patients were noted as having died. The con-

TABLE 3. Risk calculations and mortality rates for four risk algorithms

All patients (638)	STS-PROM	LES	AES	ARS
Overall mean/minimum–maximum	4.26% 0.91–46.82	13.21% 1.51–93.32	7.46% 2–22	6.99% 0.4–39.7
High-risk groups	STS-PROM (n = 64)	LES (n = 64)	AES (n = 84)	ARS (n = 97)
90th percentile risk break point	8.38%	33.47%	12%	14.3%
Expected (predicted) mortality in top 10% (mean/ minimum–maximum)	13.31% 8.38%–46.8%	50.87% 33.47%–93.32%	14.04% 12.0%–22.0%	19.03% 14.3%–39.7%
Observed perioperative mortality	18.75% (12/64)	15.63% (10/64)	11.90% (10/84)	13.40% (13/97)
Observed/expected (perioperative) mortality	1.41	0.31	0.85	0.71
Observed late mortality	45.3% (29/64)	29.7% (19/64)	33.3% (28/84)	26.8% (26/97)
Observed overall mortality	64.06% (41/64)	45.31% (29/64)	45.24% (38/84)	40.21% (39/97)
Kaplan–Meier survival estimate at 5 y	38.5% ± 6.7%	53.7% ± 7.2%	49.8% ± 6.6%	55.1% ± 6.2%
Overall mortality odds ratio,	8.06	3.27	3.45	2.76
High/low risk group	(4.63–14.01)	(1.92–5.57)	(2.14–5.57)	(1.75–4.36)

STS-PROM, The Society of Thoracic Surgeons Predicted Risk of Mortality; LES, EuroSCORE logistic; AES, EuroSCORE additive; ARS, Ambler Risk Score; CABG, coronary artery bypass grafting; NYHA, New York Heart Association.

Figure 1. Percentage of common patients within the 90th percentile of risk by model.

cordant group showed an operative mortality of 27.6% (8/29) and long-term mortality of 34.5% (10/29) for an overall rate of 62.1%. Overall mortality (operative plus follow-up deaths) was greatest in the group designated as high risk by the STS-PROM.

Logistic regression analysis demonstrated that the STS-PROM had the highest odds ratio for predicting increased overall mortality: 8.06 (95% confidence interval [CI]: 4.63–14.01) compared with 3.27 (95% CI: 1.92–5.57) for LES, 3.45 (95% CI: 2.14–5.57) for AES, and 2.76 (95% CI: 1.75–4.36) for ARS when patients were identified as high risk. A Kaplan–Meier survival curve for the four different scoring models illustrates that despite similar operative mortality among the groups, the rate of decline for the STS-PROM cohort, especially in the first year after the operation, indicates a greater sensitivity in identifying patients with the worst long-term survival than the other algorithms (Figure 2).

Discussion

The use of risk models to standardize the prediction of operative outcomes for patients undergoing cardiac surgery has gained enthusiastic support over the past several years. These tools can provide an objective assessment of potential mortality and morbidity for individual patients on the basis of specific preoperative variables and aid in gaining informed consent by presenting independent corroboration of the consulting surgeon's opinion concerning operative risk.

Several studies have demonstrated the accuracy of these models for the prediction of operative mortality.^{11–13} Increasingly, these models are also being used to identify appropriate patients for nonstandard or experimental therapy when their calculated risk profile is considered prohibitive with conventional approaches. The appeal of using these models is the ability to standardize inclusion into trials across many study sites, especially where surgical judgment comes into play and may vary between centers. The STS-PROM, EuroSCORE, and other risk models are currently being incorporated into various pivotal and feasibility studies as part of their patient selection algorithms and inclusion criteria. The usefulness of this approach depends on the accuracy of these algorithms to correctly identify the highest-risk patient populations. It is important to note that model accuracy can fluctuate over time owing to changes in surgical risk, especially for high-risk patients, and requires periodic review.

In the case of transcatheter valve implantation, the risk models should reliably identify patients at excessive risk for conventional AVR and the predicted mortality in that cohort should reasonably correlate with observed outcomes.

We compared the three most commonly used risk algorithms, STS-PROM, LES, and AES, as well as a recently published risk model (ARS) developed specifically for patients undergoing valvular heart surgery, for predictive ability in the highest risk decile of a cohort having isolated AVR for critical AS. A different number of patients out of

Figure 2. Kaplan-Meier survival curves for patients in high-risk groups as identified by different risk algorithms.

the total study population were calculated to be in the top 10% of risk depending on the model used, and a sizeable number of patients did not overlap between the models. Statistically, models demonstrate the most accuracy when a significant percentage of the study population has a particular risk factor that is included in the risk calculation. However, between each risk algorithm there can exist differences in definitions and differences in actual variables included in the risk calculation, leading to different patient stratification. For example, neurologic dysfunction in the EuroSCORE algorithm was defined to be disease that severely affects ambulation or day-to-day function, whereas the STS-PROM records only the history and timing of a previous cerebrovascular accident. The Ambler score does not include neurologic dysfunction or previous cerebrovascular accident in its calculation of risk of mortality. Other notable differences between algorithms involve the inclusion of New York Heart Association class, presence of cardiogenic shock, and need for preoperative intra-aortic balloon pump in the STS-PROM but not the AES, LES, or the ARS. Interestingly, despite the fact that more than 50 preoperative variables are collected by the STS-PROM, only 24 (listed in Table 1) are actually used in its mortality algorithm for patients having valve surgery, and preoperative ejection fraction is not included in the risk model. All models include patient age, sex, and a measure of renal function through the use of serum creatinine.

The STS-PROM underpredicted but closely approached the actual observed mortality from its identified patient group, with an observed/predicted ratio of 1.41. In contrast, the other three models overestimated predicted mortality, and all had an observed/predicted mortality of

less than 1. Notably, the concordant population identified as high risk by all of the algorithms demonstrated not only the highest operative mortality (27.6%) but also the worst long-term survival, with an additional 10 (34.5%) of 29 dying during the follow-up period.

This illustrates the potential advantage of using multiple risk algorithms with a variety of risk factors to identify a high-risk population. One hundred percent accuracy is an unrealistic expectation for any single risk model for a number of reasons. These models are developed using large heterogeneous populations of patients and are derived by weighting different risk factors on the basis of the surgical outcomes of the whole study cohort. At the highest risk level, most of the patients have the same risk profiles (Table 2), therefore providing minimal interpatient variability in which to determine risk. This limitation in risk calculation may be a factor accounting for the difference between actual risk and that predicted by each individual model. In fact, several studies have shown that both the LES and AES overestimate the mortality in octogenarians undergoing valvular heart surgery.^{14,15}

A further shortcoming of the models is that there are a number of demographic variables not collected by the risk models that have been shown to affect operative mortality.¹⁶ Hypoalbuminemia and poor nutritional status have both been shown to be independent risk factors associated with poor outcomes after cardiac surgery.¹⁷ In addition, Chang and colleagues¹⁸ have shown that patients receiving chest wall irradiation for an intrathoracic malignant tumor have reduced short- and long-term survival after cardiac surgery. Finally, frailty and debilitation are difficult conditions to objectively quantify for risk modeling and are best viewed

through the prism of surgical judgment. It is worth emphasizing that all risk models are derived from patients who have actually undergone surgery; therefore, translating these algorithms to populations previously never referred for AVR may also limit their accuracy. Ultimately, the simultaneous use of multiple models to generate a concordant risk assessment will likely surpass the efficacy of any individual model.

In terms of selecting patients for experimental therapies, some underestimation of risk represents a more conservative approach than does overprediction. Overestimation runs the risk of potentially recruiting patients for a treatment of unknown durability who might do well with conventional surgery with known efficacy and long-term results. Operative mortality has historically been the most important performance parameter collected for patients having cardiac surgery. It is an objective variable with little room for ambiguity in data collection. However, for the elderly undergoing AVR, improvement in quality of life becomes the primary goal of surgical intervention as opposed to extending a diminished life expectancy. Transcatheter valve implantation has the potential to decrease the morbidity associated with standard AVR in a high-risk population owing to the avoidance of a median sternotomy, cardiopulmonary bypass, and cardioplegic arrest. Nevertheless, the selection process necessitates thoughtful consideration of the short- and long-term risks and benefits of the procedure and a comparison of these factors with alternative therapies. Overreliance on individual risk algorithms may introduce bias owing to the limitations of the model.

In summary, the STS-PROM appears to be the most reliable single risk scoring model of those evaluated for both perioperative mortality and long-term survival after isolated AVR for AS in extremely high-risk patients. Patient identification as high risk can differ widely between different models on the basis of the variables collected and the weighting used to calculate predicted mortality. Overestimation of risk, especially when used to identify patients for experimental treatment, may lead to inclusion of patients who could do well with conventional approaches. Determining the appropriate break point in predicted risk that confers candidacy for alternative approaches can be problematic given the illustrated limitations of these four commonly used algorithms. Surgical judgment should continue to be used in conjunction with risk scoring to identify these patients.

References

1. US National Library of Medicine/National Institute of health MedlinePlus: Aortic stenosis. Available at: <http://www.nlm.nih.gov/medlineplus/ency/article/000178.htm>. Accessed April 5, 2007.
2. US Bureau of the Census. Statistical Abstract of the United States/2006. 126th ed. Washington (DC): Department of Commerce; 2006.

3. Lindroos M, Kupari M, Heikkilä J, Tilvis R. Prevalence of aortic valve abnormalities in the elderly: an echocardiographic study of a random population sample. *J Am Coll Cardiol*. 1993;21:1220-5.
4. Stewart BF, Siscovick D, Lind BK, Gardin JM, Gottdiener JS, Smith VE, et al. Clinical factors associated with calcific aortic valve disease. *J Am Coll Cardiol*. 1997;29:630-4.
5. Bonow RO, Carabello BA, Kanu C, de Leon AC Jr, Faxon DP, Freed MD, et al. ACC/AHA 2006 guidelines for the management of patients with valvular heart disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Develop Guidelines for the Management of Patients with Valvular Heart Disease). *Circulation*. 2006;114:e84-231.
6. Varadarajan P, Kapoor N, Bansal RC, Pai RG. Clinical profile and natural history of 453 nonsurgically managed patients with severe aortic stenosis. *Ann Thorac Surg*. 2006;82:2111-5.
7. Lung B, Baron G, Butchart EG, Delahaye F, Gohlke-Bärwolf C, Levang OW, et al. A prospective survey of patients with valvular heart disease in Europe: The Euro Heart Survey on Valvular Heart Disease. *Eur Heart J*. 2003;24:1231-43.
8. Webb JG, Chandavimol M, Thompson CR, Ricci DR, Carere RG, Munt BI, et al. Percutaneous aortic valve implantation retrograde from the femoral artery. *Circulation*. 2006;113:842-50.
9. Ambler G, Omar RZ, Royston P, Kinsman R, Keogh BE, Taylor KM. Generic, simple risk stratification model for heart valve surgery. *Circulation*. 2005;112:224-31.
10. Cowper DC, Kubal JD, Maynard C, Hynes DM. A primer and comparative review of major US mortality databases. *Ann Epidemiol*. 2002;12:462-8.
11. Roques F, Nashef SA, Michel P, EuroSCORE study group. Risk factors for early mortality after valve surgery in Europe in the 1990s: lessons from the EuroSCORE pilot program. *J Heart Valve Dis*. 2001;10:572-7.
12. Geissler HJ, Hölzl P, Marohl S, Kuhn-Régner F, Mehlhorn U, Südkamp M, et al. Risk stratification in heart surgery: comparison of six score systems. *Eur J Cardiothorac Surg*. 2000;17:400-6.
13. Nashef SA, Roques F, Hammill BG, Peterson ED, Michel P, Grover FL, et al. EurpSCORE Project Group. Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American cardiac surgery. *Eur J Cardiothorac Surg*. 2002;22:101-5.
14. Collart F, Feier H, Kerbaul F, Mouly-Bandini A, Riberi A, Mesana TG, et al. Valvular surgery in octogenarians: operative risk factors, evaluation of EuroSCORE and long term results. *Eur J Cardiothorac Surg*. 2005;27:276-80.
15. Collart F, Feier H, Kerbaul F, Mouly-Bandini A, Riberi A, Di Stephano E, et al. Primary valvular surgery in octogenarians: perioperative outcome. *J Heart Valve Dis*. 2005;14:238-42.
16. Higgins TL. Quantifying risk and assessing outcome in cardiac surgery. *J Cardiothorac Vasc Anesth*. 1998;12[sb:volume-nr330-340]
17. Engleman DT, Adams DH, Byrne JG, Aranki SF, Collins JJ Jr, Couper GS, et al. Impact of body mass index and albumin on morbidity and mortality after cardiac surgery. *J Thorac Cardiovasc Surg*. 1999;118:866-73.
18. Chang A, Smedira N, Chang C, Benavides MM, Myhre U, Feng J, et al. Cardiac surgery after mediastinal radiation: extent of exposure influences outcomes. *J Thorac Cardiovasc Surg*. 2007;133:404-13.

Discussion

Dr Eugene H. Blackstone (Cleveland, Ohio). Dr Dewey, I congratulate you and your colleagues on this particularly timely study. Decisions are being made, at least temporarily, about appropriate candidates for early trials of percutaneous AVR. Further, survival after device placement is being compared with that expected were open surgery performed using some of these algorithms. Your study simultaneously addresses the issue of risk stratification to identify the highest risk patients and risk calibration to quantify expected risk. You paint a somewhat disturbing picture.

As an aside, I particularly appreciate your recognizing that early risk does not stop at hospital discharge or even at 30 days, but extends further—in patients with complex problems sometimes many months—a fact that none of these models accounts for explicitly.

Your study is in the same vein as the one recently presented by the Heidelberg group at the European Association. It augments it, however, by looking at three algorithms for risk stratification and expected survival rather than just at the EuroSCORE. Let me highlight three findings of that group, because they lead to a question about your study.

First, in the highest-risk patients, EuroSCORE greatly overestimated risk, just as you have found. This means that expected risk is not well calibrated to actual risk and calls into question not just which patient is at high risk but also any attempt to use estimated risk as a barometer to assess observed risk of alternative procedures.

Second, they found that calibration progressively worsened with era of operation. The EuroSCORE was based on data from 1995; 10 years later, it should not be expected to be well calibrated to contemporary results. It needs to be updated, as do some other measures.

But third, they nevertheless found that as a tool to rank patients according to risk, that is, risk stratification, performance remained good.

So question one is, have you examined performance of these scores across your experience in terms of both calibration to actual risk and consistency in ranking patients?

Now, a different sort of observation I would like to make is related to what today I would call prediction error, although in earlier times I called it residual risk. When we examined prediction error for time-related survival of nearly 4000 Belgian patients after coronary artery bypass grafting, we found three categories of error: (1) variables available but ignored that actually were associated with risk, which is one of the points you make in your discussion; (2) variables representing rare subgroups of patients not present in sufficient abundance at time of model development to estimate risk reliably; and (3) variables representing subgroups completely absent from the mix of patients on whom the model was developed.

It seems that you have identified all three of these pitfalls in your study: ignored variables, variables related to unusual complexity of frail, elderly patients being considered for alternative therapy, and a subgroup of patients not represented in the mix of open surgery patients for whom these models were developed. This leads me to a second question followed by an observation: What was the survival of patients who were identified concordantly by all three algorithms and, as a corollary, what was the survival of the discordant patients? I ask this to encourage you not to be terribly discouraged by somewhat discordant results by three or four different models. Indeed, I believe this points to the correct direction for the future.

Here is what is going on in analysis of genetic data and a number of other fields. Rather than developing a single model such as a classification or risk stratification tree, thousands and even tens of thousands of models are built from the data. Building from a single classification tree to multiple forests of trees (based completely on minimizing prediction error) may pave the way for best identifying high-risk patients and possibly even best predict-

ing their expected survival. But it will work only if such patients are already present in our databases or if we become really good, as we are not now, at extrapolating beyond our data.

Dr Dewey. Thank you very much, Dr Blackstone, for your kind comments and for taking your time to review our manuscript. As usual, you find a unique way of looking at the data that others of us have not quite thought about, and I think your comments are actually very appropriate and cogent.

We have not examined the performance of these scores across our entire experience, and we have not specifically looked at how the patients identified by each of the algorithms, how their survival compares with the ones who were not discordant within the algorithms. One would assume that the patients in common, which ranged to about 60% for the most part, are the ones that are dying, and the reasons that there are differences either in perioperative mortality or long-term follow-up is that the other risk algorithms are diluted by other patients who do better than these truly higher-risk patients.

Your comments in regard to updating and continually reviewing these risk algorithms are particularly appropriate. The STS algorithm patient set was defined from 1994 to 1997, and our patient population may have actually changed over that time period. That is a decade ago and probably need to be updated and looked at to see how valid it is during this decade.

One of the tensions with these algorithms is that these are designed to be used at the bedside as a simple sort of tool that surgeons can use, so that when you go into the patient's room and talk to the patient, you have some numbers available other than your experience or your gestalt. And while not ever going to be as accurate as if you use multiple different algorithms based on much larger data sets and variables, I think they are still relatively useful for the tools in which we use them.

Dr Robert A. Guyton (*Atlanta, Ga*). That is a very nice study. A recent presentation of Core Valve patients, 22 patients from Germany, presented an LES score averaging 11%. When we used the STS prediction to roughly estimate this predictive mortality, it was more in the range of 6%. The LES data appeared to overestimate the risk, and these were patients undergoing percutaneous valve procedures, with a 6% to 10% operative mortality based on these predictions. On the basis of your study, would it be appropriate to recommend that patients being included in percutaneous valve studies be assessed by more than one, perhaps at least three, of the existing prediction systems so that we have a good idea of what range of operative mortality that we are talking about in these patients who are now undergoing percutaneous or transapical therapy?

Dr Dewey. That is a very good comment, Dr Guyton. I think I would be very careful about relying just on the LES, because I think not only our study but also other studies show that in high-risk patient populations it tends to overestimate risk. Now, I would argue that an LES of 11 is not a high-risk population to begin with. One of the areas that we have in the percutaneous valve patients, however, is these are typically patients who have not had surgery, theoretically, and so how you incorporate that patient set into a pool of patients who had surgery in which these algorithms were developed is somewhat tricky. Our risk algorithms may not be nearly as applicable to those patients who might never have had surgery as some of the patients that we are looking at currently.