Case complexity scores in congenital heart surgery: A comparative study of the Aristotle Basic Complexity score and the Risk Adjustment in Congenital Heart Surgery (RACHS-1) system

Osman O. Al-Radi, MD, MSc,[a] Frank E. Harrell Jr, PhD,[b] Christopher A. Caldarone, MD,[a] Brian W. McCrindle, MD, MPH,[a] Jeffrey P. Jacobs, MD,[c] M. Gail Williams,[a] Glen S. Van Arsdell, MD,[a] and William G. Williams, MD[a]

Dr Al-Radi

From The Hospital for Sick Children,[a] University of Toronto, Toronto, Canada; Department of Biostatistics,[b] Vanderbilt University School of Medicine, Nashville, Tenn.; and The Congenital Heart Institute of Florida,[c] University of South Florida, Saint Petersburg, Fla.

Read at the Eighty-fifth Annual Meeting of The American Association for Thoracic Surgery, San Francisco, Calif, April 10-13, 2005.

Received for publication April 20, 2005; revisions received April 26, 2006; accepted for publication May 17, 2006.

Address for reprints: William G. Williams, MD, 555 University Avenue, Room 1525, Toronto, ON, M5G 1X8, Canada (E-mail: bill.williams@sickkids.ca).

J Thorac Cardiovasc Surg 2007;133:865-74

0022-5223/$32.00

doi:10.1016/j.jtcvs.2006.05.071

Objective: The Aristotle Basic Complexity score and the Risk Adjustment in Congenital Heart Surgery system were developed by consensus to compare outcomes of congenital cardiac surgery. We compared the predictive value of the 2 systems.

Methods: Of all index congenital cardiac operations at our institution from 1982 to 2004 (n = 13,675), we were able to assign an Aristotle Basic Complexity score, a Risk Adjustment in Congenital Heart Surgery score, and both scores to 13,138 (96%), 11,533 (84%), and 11,438 (84%) operations, respectively. Models of in-hospital mortality and length of stay were generated for Aristotle Basic Complexity and Risk Adjustment in Congenital Heart Surgery using an identical data set in which both Aristotle Basic Complexity and Risk Adjustment in Congenital Heart Surgery scores were assigned. The likelihood ratio test for nested models and paired concordance statistics were used.

Results: After adjustment for year of operation, the odds ratios for Aristotle Basic Complexity score 3 versus 6, 9 versus 6, 12 versus 6, and 15 versus 6 were 0.29, 2.22, 7.62, and 26.54 ($P < .0001$). Similarly, odds ratios for Risk Adjustment in Congenital Heart Surgery categories 1 versus 2, 3 versus 2, 4 versus 2, and 5/6 versus 2 were 0.23, 1.98, 5.80, and 20.71 ($P < .0001$). Risk Adjustment in Congenital Heart Surgery added significant predictive value over Aristotle Basic Complexity (likelihood ratio $\chi^2 = 162$, $P < .0001$), whereas Aristotle Basic Complexity contributed much less predictive value over Risk Adjustment in Congenital Heart Surgery (likelihood ratio $\chi^2 = 13.4$, $P = .009$). Neither system fully adjusted for the child's age. The Risk Adjustment in Congenital Heart Surgery scores were more concordant with length of stay compared with Aristotle Basic Complexity scores ($P < .0001$).

Conclusions: The predictive value of Risk Adjustment in Congenital Heart Surgery is higher than that of Aristotle Basic Complexity. The use of Aristotle Basic Complexity or Risk Adjustment in Congenital Heart Surgery as risk stratification and trending tools to monitor outcomes over time and to guide risk-adjusted comparisons may be valuable.

Because each congenital heart defect is a rare condition, assessing the quality of care based on crude outcomes is problematic. Several groups have proposed systems of assessing quality of care by assigning surgical operations a risk score or grouping operations of similar risk into categories. Two such systems, namely the Aristotle Basic Complexity (ABC) score and the Risk Adjustment for Congenital Heart Surgery (RACHS-1), were developed by consensus of

CHD

Abbreviations and Acronyms

ABC = Aristotle Basic Complexity
CI = confidence interval
CVSDB = cardiovascular surgery database
LR = likelihood ratio
OR = odds ratio
RACHS-1 = Risk Adjustment in Congenital Heart Surgery
ROC = receiver operator characteristics

experts.[1,2] The methodologic details of each system are described in the respective references. Briefly, the Aristotle committee, consisting of experts from 50 centers in 23 countries, developed the ABC score. Potential for mortality, potential for morbidity, and technical difficulty for each operation contribute up to 5 points each to this continuous score (range 1.5 to 15). The score was used by its authors to group the procedures as follows: level 1, scores 1.5 to 5.9; level 2, scores 6 to 7.9; level 3, scores 8 to 9.9; and level 4, scores 10 to 15.[3-5] On the other hand, in the RACHS-1 system, which was developed between 1993 and 1995, congenital cardiac operations were stratified into 1 of 6 categories. The risk category of some procedures additionally varied depending on age.[6-7] RACHS-1 was validated in 2 independent populations and was found to have good predictive value.[8,9] However, no studies have compared the predictive value of the 2 systems in the same population.

We sought to assess the predictive value of ABC and RACHS-1 by comparing in-hospital mortality and length of stay as predicted by the respective system with the observed in-hospital mortality and length of stay at our institution.[10]

Materials and Methods

The outcomes of 13,675 index (first operation of an admission) congenital cardiac surgeries performed on children (age $<$ 18 years) between July 1, 1982 and June 30, 2004 were available in the cardiovascular surgery database (CVSDB) at the Hospital for Sick Children. These index operations were performed on 10,860 children who had a total of 16,538 cardiac operations. More than 1 index operation was performed on 1937 (19% of 10,860) children. Only the index operations with both ABC score and RACHS-1 category assigned were included in the study (n = 11,438). CVSDB is a prospective clinical database. Data in CVSDB are maintained by a dedicated staff and are validated by monthly audits of operating room logs, surgeons' office files, morbidity and mortality conferences, and a clinical nurse coordinator. Monthly output reports from CVSDB are sent to the faculty. An automated algorithm was used to assign ABC score and RACHS-1 values to the procedure codes in CVSDB. Because of the paucity of RACHS-1 category 5, it was combined with category 6.

For in-hospital mortality, logistic regression models were generated for ABC and RACHS-1 separately and then combined in 1 model. All 3 models included an identical set of operations. ABC score was modeled as a continuous variable with appropriate transformation using restricted cubic splines to account for nonlinear relationships.[11] Compared with traditional transformations (log, square root, or polynomials), cubic splines are better suited for biologic associations as they allow flexibility for nonlinear relationships.[12] The locations of change in the curvature are set at specific points known as knots. Five knots at quantiles of the predictors were used in our analyses. Because 1937 children had more than 1 index operation, Huber–White robust sandwich estimates of the variance covariance matrix were used to penalize for clustering by patient in all logistic models.[13-16] The predictive value of the models was assessed by the area under the receiver operator characteristics (ROC) curve, also known as the c-index, the model likelihood ratio (LR) χ^2 statistic, and the adequacy index.[17] To test for a difference in the predictive value of the 2 systems, we used the LR χ^2 test for nested models to assess whether ABC adds predictive value to a model that includes RACHS-1 and whether RACHS-1 adds predictive value to a model that includes ABC. These analyses were done with and without adjustment for year of operation and the child's age at operation. Such tests are more sensitive than tests comparing ROC areas (c-index).[11] However, the comparison between the ROC areas was also done and presented for the sake of completeness. The latter was obtained using bootstrap confidence intervals (CIs) from 1000 resamples. The adequacy index is the fraction of the total LR χ^2 explained by a set of variables that could be explained by omitting the competing variable. The clinical utility of predictive models was assessed by the frequency of patients identified by the model with very low or very high risk of death. Models with higher frequency of extreme predictions are more likely to be clinically useful. Model calibration was assessed by bootstrap estimates of predicted mortality versus actual mortality.[11]

For length of stay, a rank correlation U-statistic for paired censored data was used to estimate the fraction of pairs for which the prediction using RACHS-1 was more discriminating compared with ABC,[14] and both were analyzed as continuous variables. Patients who died before discharge were censored in this analysis. A competing risk analysis without censoring death but rather treating it as a competing event was also conducted to produce cumulative incidence plots.[18,19]

Mathematical representations of the logistic models are presented in an appendix (Appendix A). The R statistical package, Hmisc,[14] Design,[11] and Cmprsk[18,19] libraries (www.r-project.org) were used for all analyses.

Results

Of the 13,675 index operations in CVSDB, an ABC score could be assigned to 13,138 (96%) operations, a RACHS-1 category could be assigned to 11,533 (84%) operations, and both ABC and RACHS-1 could be assigned to 11,438 (84%) operations. Only operations that were assigned both ABC and RACHS-1 (n = 11,438) were used in all subsequent analyses. Patient demographics and crude outcomes are presented in Table 1. Exploratory plots of hospital death (observed and predicted) versus ABC score levels and RACHS-1 categories are shown in Figure 1, A and B,

TABLE 1. Patient demographics and outcomes by surgical era and risk group

	ABC				RACHS-1				
	1 (n = 1771)	2 (n = 5238)	3 (n = 3106)	4 (n = 1323)	1 (n = 1965)	2 (n = 4365)	3 (n = 3873)	4 (n = 925)	5/6 (n = 310)
1982-1988									
n	593	1436	763	183	635	979	1070	268	23
Age (mo)	50.2	46.8	50.2	26.2	55.7	48.8	47.7	21.4	4.9
Weight (kg)	15.7	15.0	14.8	9.1	17.5	15.1	14.5	8.6	4.1
Mortality	0.03	0.06	0.11	0.31	0.01	0.06	0.09	0.27	0.78
LOS (d)	7	10	12	12	7	10	12	13	1
1988-1993									
n	389	1348	768	326	445	1119	980	229	58
Age (mo)	56.4	40.7	40.0	10.3	55.7	37.3	40.1	21.9	0.4
Weight (kg)	17.4	14.1	12.9	5.6	17.6	13	13.1	9.1	3.3
Mortality	0.03	0.07	0.09	0.27	0.01	0.05	0.10	0.22	0.66
LOS (d)	7	9	13.5	13	7	10	13	13	4
1993-1999									
n	519	1363	898	455	580	1256	1040	240	119
Age (mo)	59.5	36.5	27.8	5.4	55.2	29.7	32.5	19.1	2.7
Weight (kg)	18.4	13.5	10.9	4.7	17.6	11.7	11.8	9.1	4
Mortality	0.01	0.05	0.07	0.21	0.01	0.04	0.08	0.2	0.4
LOS (d)	4	7	10	12	4	8	10	11	15
1999-2004									
n	270	1091	677	359	305	1011	783	188	110
Age (mo)	42.6	38.8	27.4	4.2	40.3	29.3	34.6	25.5	0.7
Weight (kg)	14.2	14.1	11.3	4.5	13.6	11.8	12.8	10.6	3.4
Mortality	0	0.02	0.03	0.07	0	0.01	0.04	0.07	0.17
LOS (d)	3	7	10	15	3	7	10	12	27

Age and weight are presented as means. Mortality: Fraction of patients who died before hospital discharge. *LOS*, Median length of stay in hospital after surgery. ABC score was divided into levels: 1, scores 1.5 to 5.9; 2, 6 to 7.9; 3, 8 to 9.9; and 4, 10 to 15.

respectively. As illustrated in both plots, the probability of hospital death declined with time and markedly so during the 1990s. Improvement in survival was most significant for high-risk procedures, such as those with ABC score above 9.9 or RACHS-1 categories 4 and 5/6 (Figure 1).

Is ABC and/or RACHS-1 Predictive of In-hospital Mortality?

Both ABC and RACHS-1 were predictive of in-hospital mortality. ABC score 6, which was the median score, was chosen as the reference score. The odds ratios (ORs) adjusted for year of operation were 0.29, 2.22, 7.62, and 26.54 for ABC scores 3 versus 6, 9 versus 6, 12 versus 6, and 15 versus 6 (95% CIs: 0.18-0.46, 1.83-2.68, 6.21-9.34, and 19.32-36.45, respectively; $P < .0001$ overall and .02 for the nonlinear component).

Similarly, RACHS-1 category 2, which was the median category, was chosen as the reference category. The ORs adjusted for year of operation were 0.23, 1.98, 5.80, and 20.71 for RACHS-1 categories 1 versus 2, 3 versus 2, 4 versus 2, and 5/6 versus 2 (CIs: 0.14-0.36, 1.64-2.40, 4.64-7.26, and 15.52-27.64, respectively; $P < .0001$).

Is the Predictive Value of ABC Higher or Lower than That of RACHS-1?

The predictive values measured by the c-index and LR χ^2 for ABC score and RACHS-1 with and without adjustment for the year of operation are shown in Table 2. Both the c-index and the LR χ^2 are higher for RACHS-1 models. Using the LR χ^2 test for nested models, ABC did not add predictive value to a model that includes RACHS-1 (LR $\chi^2 = 6.2$, df = 4, $P = .18$), whereas RACHS-1 added clinically and statistically significant predictive value to a model that includes ABC (LR $\chi^2 = 182$, df = 4, $P < .0001$). The difference between the c-index of ABC and RACHS-1 models was also significant ($P = .018$, c-index 0.698 vs 0.733, respectively).

After adjustment for year of operation, however, ABC added a clinically small but statistically significant predictive value to RACHS-1 (LR $\chi^2 = 13.4$, df = 4, $P = .009$), whereas RACHS-1 continued to add clinically and statistically significant predictive value to ABC (LR $\chi^2 = 162$, df = 4, $P < .0001$). The difference between the c-index of ABC and RACHS-1 models adjusted for year of operation was also significant ($P = .03$, c-index 0.737 vs 0.763, respectively).

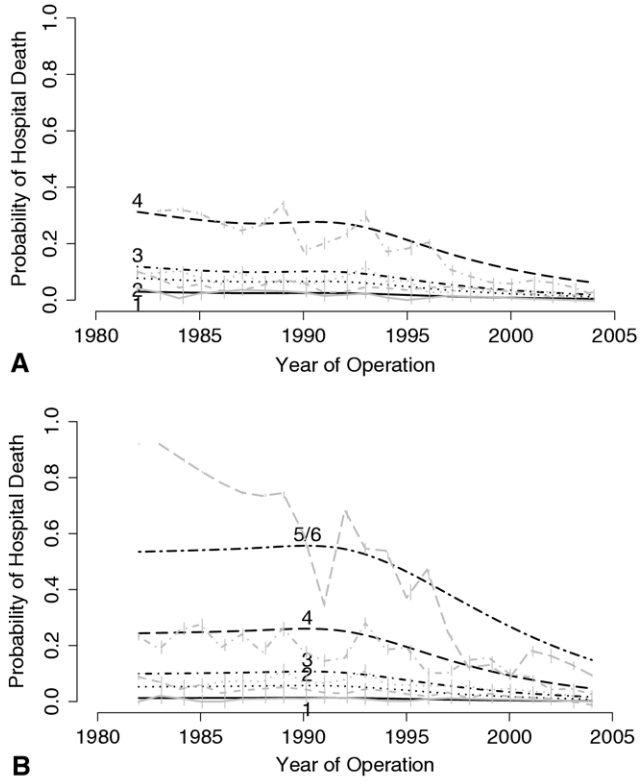

Figure 1. Observed *(gray)* and predicted *(black)* probability of in-hospital death over calendar year of operation. A, For ABC levels (score range), 1 (1.5-5.9), 2 (6-7.9), 3 (8-9.9), and 4 (10-15). B, For RACHS-1 categories 1, 2, 3, 4, and 5/6.

The adequacy index (ie, the proportion of predictive LR χ^2 value attributable to ABC score) and ABC score adjusted for year of operation were 72% and 80%, respectively. On the other hand, the adequacy index for RACHS-1 and RACHS-1 adjusted for year of operation was 99% and 98%, respectively. Therefore, ABC was sensitive to adjustment for the year of operation, whereas RACHS-1 was not. Figure 2 summarizes the comparison between the predictive value of ABC and that of RACHS-1, as well as the effect of adjustments for year of operation.

Is the Difference Between ABC and RACHS-1 of Clinical Importance?

In addition to the predictive power, the ability ABC versus RACHS-1 to identify patients with very low (<1%) or very high (>15%) risk of in-hospital death was different. As seen in Figure 3, RACHS-1 identified more children with <1% or >15% risk of in-hospital death compared with ABC. When both ABC and RACHS-1 were combined, the extreme predictions were even higher, suggesting additive information from ABC and RACHS-1.

Does the Addition of Other Predictive Factors Improve the Predictions Made by Either ABC or RACHS-1?

Age at the operation was an important predictive factor in predicting in-hospital mortality ($\chi^2 = 155$, $P < .0001$). When age in months was added as a new predictive factor to a model that included ABC score, the predictive value of such a model significantly improved (LR $\chi^2 = 366$, df = 4, $P < .0001$). Similarly, but to a lesser extent, adding age to a model that included RACHS-1 (which intrinsically partially adjusts for age) also improved the predictive value of such a model (LR $\chi^2 = 226$, df = 4, $P < .0001$). Neither ABC nor RACHS-1 adjusts adequately for the child's age at operation. Furthermore, the effect of combining ABC and RACHS-1 was significantly different from that of either system alone. However, RACHS-1 adds much more predictive value to ABC compared with what ABC adds to RACHS-1. Using the adequacy index (see Figure 2), in models adjusted for year of operation, ABC adds 2% to RACHS-1, where as RACHS-1 adds 18% to ABC.

Is ABC Score and/or RACHS-1 Associated With the Child's Length of Stay in the Hospital?

Postoperative length of stay in the hospital was strongly associated with year of operation, ABC score, and RACHS-1. RACHS-1, however, was more concordant with length of stay compared with ABC ($P < .0001$). A competing risk analysis demonstrated that both ABC and RACHS-1 were predictive of hospital discharge and death when they were treated as competing risk events. The cumulative incidence plots for death and discharge from hospital for each ABC level or RACHS-1 category are presented in Figure 4, A and B, respectively. The risk of death increased with each increase in ABC score or

TABLE 2. Model discrimination statistics for logistic models of in-hospital death

Model	Degrees of freedom	C-index			LR χ^2	
		ABC	RACHS-1	ABC vs RACHS-1, *P* value	ABC	RACHS-1
Unadjusted	4	0.698	0.733	.018	490	667
Adjusted to year of operation	8	0.737	0.763	.03	677	828

The *P* values provided are based on a nonparametric bootstrap confidence interval.

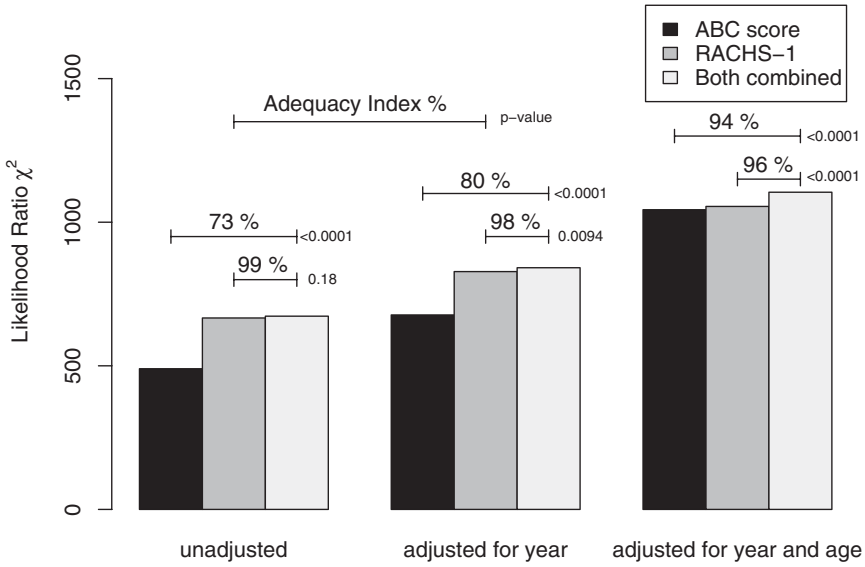

Figure 2. The predictive value of robust logistic models of ABC, RACHS-1, and both scores combined. Unadjusted models, models adjusted for the calendar year of operation, and models adjusted for calendar year of operation and the child's age at operation are presented. The predictive value is represented by the model LR and adequacy index (% likelihood ratio χ^2 attributable to the nested model, which does include both ABC and RACHS-1; see Methods). *P* values shown are of an LR χ^2 test for nested models with 4 degrees of freedom.

RACHS-1. The mean length of stay decreased for each increase in ABC score or RACHS-1.

In summary, the predictive value of RACHS-1 for in-hospital mortality and length of stay was higher than that of ABC. Adjustment for the year of operation improved the predictive value of both systems; however, a significant difference in predictive values between ABC and RACHS-1 persisted. Adjustment for the age of the child increased the predictive value of ABC score to a level very close to that of RACHS-1. There may additional gain in predictive value if ABC and RACHS-1 are combined.

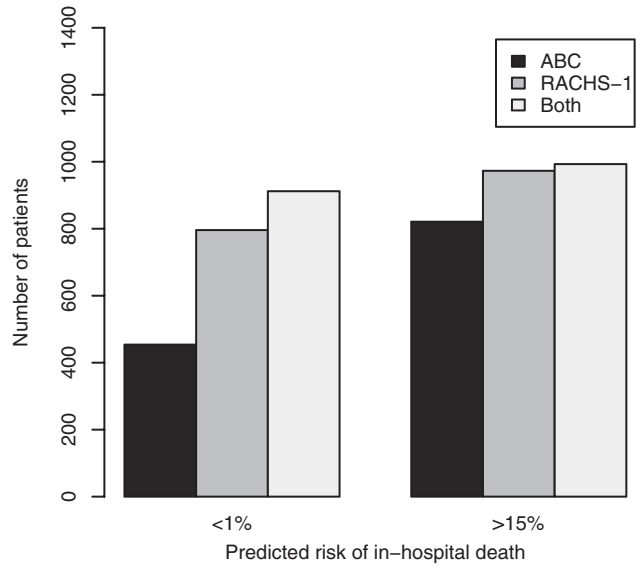

Figure 3. The number of patients *(Frequency)* assigned to intervals of predicted risk by ABC, RACHS-1, or both with extreme (ie, very low [<1%] and very high [>15%]) scores predicted in-hospital mortality. Models with more frequent extreme predictions are more clinically useful. The cutoffs of 1% and 15% were approximately the 5th and 95th percentiles of predicted risk of in-hospital death.

Discussion

RACHS-1, as the name implies, was developed as a method of risk adjustment in congenital heart surgery.[1] Our analysis showed that RACHS-1 was strongly associated with in-hospital mortality and with length of stay. However, RACHS-1 is better characterized as a method of risk stratification than as adjustment. As Jenkins and colleagues[1] acknowledged, every child is different. We have shown that the child's age at operation is an important prognostic factor that is only partially accounted for within a few RACHS-1 category codes. The predictive value of future versions of RACHS-1 may be improved by adding an adjustment for age to more diagnostic codes. The addition of other procedure-specific key predictive factors may further improve RACHS-1 predictive value.

The Aristotle committee intended to assess the "performance" of surgical care providers and hypothesized that performance = outcomes × complexity.[2] We did not assess this hypothesis; rather, we focused on assessing ABC as it correlates with short-term outcomes, namely in-hospital mortality and length of stay. There was a strong association between ABC and in-hospital mortality and length of stay; however, its predictive value was lower than that of RACHS-1. This was attributable to some extent to its failure to adjust for the child's age. When we adjusted ABC by including age at operation in predictive models with ABC, the predictive value of such models improved to a level very close to that of RACHS-1.

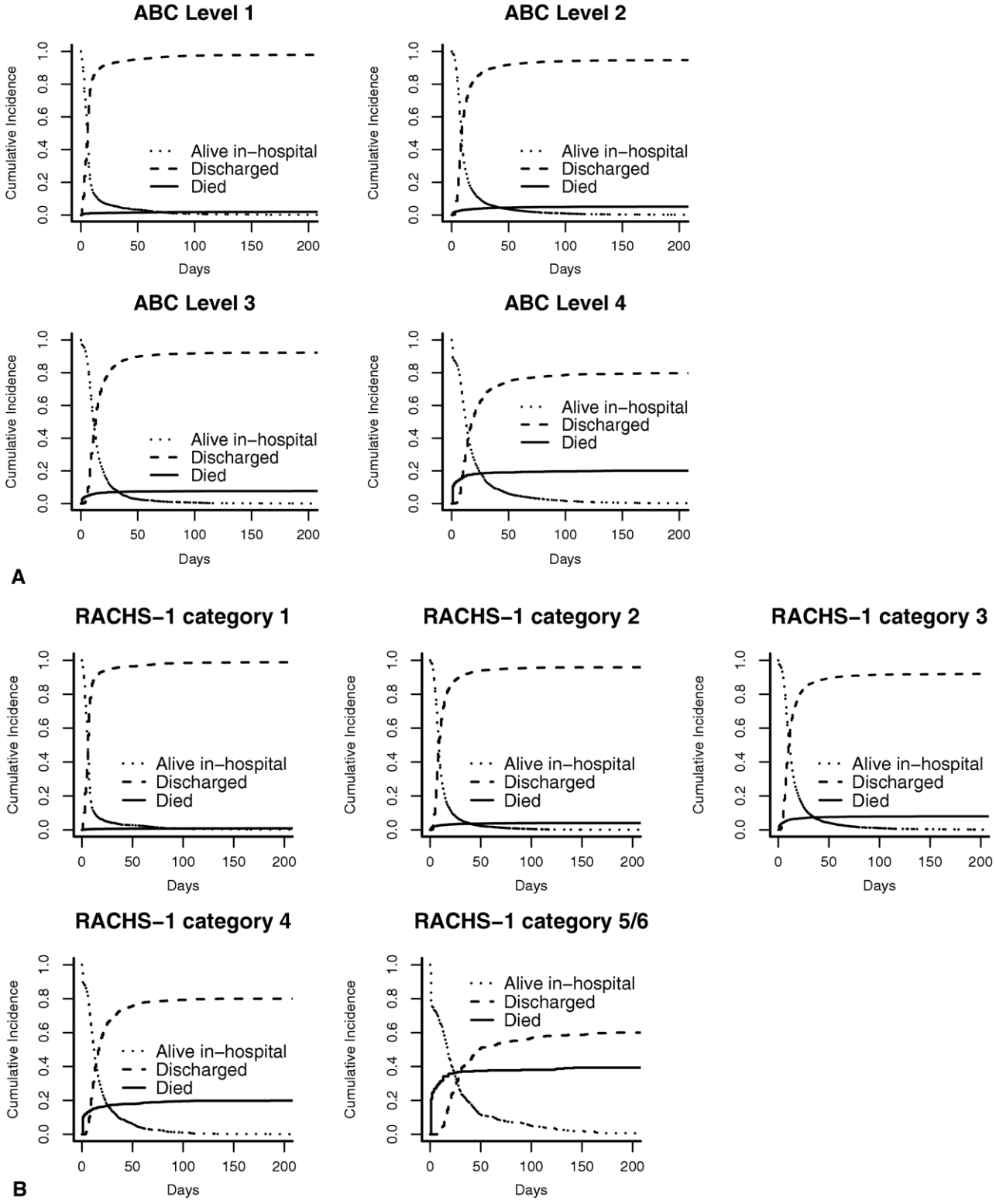

Figure 4. Cumulative incidence plots of death before discharge versus discharge alive over time from the day of operation (time 0). A, For ABC levels (score range), 1 (1.5-5.9), 2 (6-7.9), 3 (8-9.9), and 4 (10-15). B, For RACHS-1 categories 1, 2, 3, 4, and 5/6. The time point (x-axis point) corresponding to a cumulative incidence (y-axis point) of 0.5 on the *dashed curve* represents the mean time to discharge from hospital.

Further, the predictions of both systems need to be adjusted for the year of operation to account for improvements in outcomes of congenital heart surgery that have occurred over the last 2 decades and that will continue to occur in the future.

ROC curves are used to compare the predictive value of risk models extensively in the medical literature. However, because the resultant c-index (area under the ROC curve) is a rank-based statistic, it fails to reward extreme predictions that are true and fails to penalize extreme predictions that are false. Therefore, the c-index is insensitive to potentially important differences; a small difference in the c-index may actually be of large clinical and statistical importance. For this study, we chose a more sensitive statistic known as the LR χ^2, which is not rank based and appropriately rewards correct extreme predictions. In the case of our analysis, due to the large sample size, the 2 methodologies result in the same conclusions, which were both presented, namely a significant difference between ABC and RACHS-1. However, the magnitude of the difference is more prominent using the LR method versus the ROC c-index method.

Both ABC and RACHS-1 are, however, useful guides to assess the quality of surgical care providers over time. The graphical exploration of trends over time and a comparison of institutional outcomes within risk levels may be useful in detecting outliers and in generating hypotheses about differences between institutions or methods of care, as in Figures 1 and 4. Importantly, however, a comparison of institutions on the basis of in-hospital mortality is a very blunt measure of a much more complex scenario and unlikely to allow fair comparisons or meaningful information upon which to base changes in practice. Specific hypotheses to compare or improve quality of care must be tested with truly risk-adjusted models using more comprehensive data, in a way that would let the data speak for themselves.[20]

Limitations

More index operations were assigned an ABC score than were assigned a RACHS-1 category, 94% versus 86%. However, to adequately compare the 2 systems, only operations that were assigned both an ABC score and a RACHS-1 category were included in the analysis. When analyses were done using all possible procedures that were assigned either an ABC score or a RACHS-1 category, the overall conclusions were the same as those of the analyses presented here.

Because 19% of the children had more than 1 index operation, clustering of the outcome by patient was taken into account. This was achieved by penalizing the model estimates to account for clustering.

In-hospital mortality was the outcome available, and it was rigorously validated in our database. However, it does not completely represent the mortality associated with the early hazard phase described post-cardiac surgery.[21] Notwithstanding the fact that the 2 systems (ABC and RACHS-1) were designed to predict short-term outcomes, the relationship between the scores and time-related survival, both in early and late hazard phases, will be of future interest.

Conclusions

We have shown that both ABC and RACHS-1 have a strong association with in-hospital mortality and length of stay. The predictive value of RACHS-1 is higher than that of ABC. Adding patient- and procedure-specific variables may improve their predictive value. Neither system in isolation is adequate for risk-adjusted comparisons between providers of care or institutions. Their use as risk stratification and trending tools to monitor outcomes over time, with the intention to guide risk-adjusted comparisons, could be valuable.

References

1. Jenkins KJ, Gauvreau K, Newburger JW, Spray TL, Moller JH, Iezzoni LI. Consensus-based method for risk adjustment for surgery for congenital heart disease. *J Thorac Cardiovasc Surg.* 2002;123:110-8.
2. Lacour-Gayet F, Clarke D, Jacobs J, Comas J, Daebritz S, Daenen W, et al. The Aristotle score: a complexity-adjusted method to evaluate surgical results. *Eur J Cardiothorac Surg.* 2004;25:911-24.
3. Lacour-Gayet F. Risk stratification theme for congenital heart surgery. *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu.* 2002;5:148-52.
4. Lacour-Gayet F, Clarke D, Jacobs J, Gaynor W, Hamilton L, Jacobs M, et al. The Aristotle score for congenital heart surgery. *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu.* 2004;7:185-91.
5. Lacour-Gayet F, Clarke DR. The Aristotle method: a new concept to evaluate quality of care based on complexity. *Curr Opin Pediatr.* 2005;17:412-7.
6. Jenkins KJ, Gauvreau K. Center-specific differences in mortality: preliminary analyses using the Risk Adjustment in Congenital Heart Surgery (RACHS-1) method. *J Thorac Cardiovasc Surg.* 2002;124:97-104.
7. Jenkins KJ. Risk adjustment for congenital heart surgery: the RACHS-1 method. *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu.* 2004;7:180-4.
8. Larsen SH, Pedersen J, Jacobsen J, Johnsen SP, Hansen OK, Hjortdal V. The RACHS-1 risk categories reflect mortality and length of stay in a Danish population of children operated for congenital heart disease. *Eur J Cardiothorac Surg.* 2005;28:877-81.
9. Boethig D, Jenkins KJ, Hecker H, Thies WR, Breymann T. The RACHS-1 risk categories reflect mortality and length of hospital stay in a large German pediatric cardiac surgery population. *Eur J Cardiothorac Surg.* 2004;26:12-7.
10. Iezzoni LI. Risk adjustment for measuring health care outcomes. 3rd ed. Chicago: Health Administration Press; 2003.
11. Harrell FE Jr. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
12. Harrell FE Jr, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *J Natl Cancer Inst.* 1988;80:1198-202.
13. White H. Maximum likelihood estimation of misspecified models. Econometrica 1982;50:1-25.
14. Hmisc: A Package of Miscellaneous S Functions. 2006.
15. Proc Fifth Berkeley Symposium Math Stat. 1967.
16. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. *Stat Med.* 1996;15:1793-806.
17. Califf RM, Phillips HR III, Hindman MC, Mark DB, Lee KL, Behar VS, et al. Prognostic value of a coronary artery jeopardy score. *J Am Coll Cardiol.* 1985;5:1055-63.

18. Aalen OO. Nonparametric estimation of partial transition probabilities in multiple decrement models. *Annals of Statistics.* 2006;61:534-45.
19. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics.* 1988;16:1141-54.
20. Blackstone EH. Let the data speak for themselves? *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu.* 2004;7:192-8.
21. Monro JL. The next challenge-adapting to change. *Eur J Cardiothorac Surg.* 2004;26:1063-72.

CHD

Appendix

Logistic models 1, 2, and 3 were used to model ABC score, unadjusted, adjusted for year of operation, and adjusted for year of operation and the child's age, respectively. In-hospital death was the outcome variable.

Model 1

$$\text{Prob}\{\text{DEATH} = 1\} = \frac{1}{1 + \exp(-X\beta)},$$

where $X\hat{\beta} = -6.006689 + 0.5913527\ \text{ABC} - 0.01976131\ (\text{ABC-3})_+^3 + 0.2401961(\text{ABC-6})_+^3 - 0.2436142(\text{ABC-6.5})_+^3 + 0.02857559(\text{ABC-9})_+^3 - 0.005396189(\text{ABC-10.3})_+^3$ and $(x) = x$ if $x > 0$, 0 otherwise.

Model 2

$$\text{Prob}\{\text{DEATH} = 1\} = \frac{1}{1 + \exp(-X\beta)},$$

where $X\hat{\beta} = 63.52153 + 0.5333782\text{ABC} - 0.01307427(\text{ABC-3})_+^3 + 0.1365018(\text{ABC-6})_+^3 - 0.1291932(\text{ABC-6.5})_+^3 - 0.0004472514(\text{ABC-9})_+^3 + 0.006212918(\text{ABC-10.3})_+^3 - 0.03479037\ \text{Year} + 0.000424993(\text{Year-1983})_+^3 - 0.002778462(\text{Year-1988})_+^3 + 0.003802972(\text{Year-1992})_+^3 - 0.001442604(\text{Year-1997})_+^3 - 6.898965 \times 10^{-6}(\text{Year-2003})_+^3$ and $(x)_- = x$ if $x > 0$, 0 otherwise.

Model 3

$$\text{Prob}\{\text{DEATH} = 1\} = \frac{1}{1 + \exp(-X\beta)},$$

where $X\hat{\beta} = 37.21254 - 0.2004645\text{ABC} + 0.04727446(\text{ABC-3})_+^3 - 0.7220843(\text{ABC-6})_+^3 + 0.7783477(\text{ABC-6.5})_+^3 - 0.1522020(\text{ABC-9})_+^3 + 0.04866409(\text{ABC-10.3})_+^3 - 0.01958611\text{Year} - 0.000331136(\text{Year-1983})_+^3 + 4.973796 \times 10^{-5}(\text{Year-1988})_+^3 + 0.0004777783(\text{Year-1992})_+^3 + 0.000103515(\text{Year-1997})_+^3 - 0.0002998952(\text{Year-2003})_+^3 - 0.3045867\text{Age} + 0.002549708(\text{Age-0.13})_+^3 - 0.003324442(\text{Age-3.05})_+^3 + 0.0007674379(\text{Age-12.33})_+^3 + 7.35287 \times 10^{-6}\ (\text{Age-47.61})_+^3 - 5.142649 \times 10^{-8}\ (\text{Age-164.96})_+^3$ and $(x)_+ = x$ if $x > 0$, 0 otherwise.

The model coefficients and predictive value statistics are shown in Table A1. Logistic models 4, 5, and 6 were used to model RACHS-1, unadjusted, adjusted for year of operation, and adjusted for year of operation and the child's age, respectively. In-hospital death was the outcome variable:

Model 4

$$\text{Prob}\{\text{DEATH} = 1\} = \frac{1}{1 + \exp(-X\beta)},$$

where $X\hat{\beta} = -4.577 + 1.413\ \{\text{RACHS} = 2\} + 2.132\ \{\text{RACHS} = 3\} + 3.184\ \{\text{RACHS} = 4\} + 4.158\ \{\text{RACHS} = 5/6\}$ and $\{c\} = 1$ if subject is in group c, 0 otherwise; $(x)_+ = x$ if $x > 0$, 0 otherwise

Model 5

$$\text{Prob}\{\text{DEATH} = 1\} = \frac{1}{1 + \exp(-X\beta)},$$

where $X\hat{\beta} = -18.94 + 1.486\ \{\text{RACHS} = 2\} + 2.171\ \{\text{RACHS} = 3\} + 3.245\ \{\text{RACHS} = 4\} + 4.517\ \{\text{RACHS} = 5/6\} + 0.00735\ \text{Year} + 0.0001281(\text{Year-1983})_+^3 - 0.001991(\text{Year-1988})_+^3 + 0.003059(\text{Year-1992})_+^3 - 0.001056(\text{Year-1997})_+^3 - 0.0001395(\text{Year-2003})_+^3$ and $\{c\} = 1$ if subject is in group c, 0 otherwise; $(x)_+ = x$ if $x > 0$, 0 otherwise

Model 6

$$\text{Prob}\{\text{DEATH} = 1\} = \frac{1}{1 + \exp(-X\beta)},$$

where $X\hat{\beta} = -55.25 + 1.372\ \{\text{RACHS} = 2\} + 1.902\ \{\text{RACHS} = 3\} + 2.544\ \{\text{RACHS} = 4\} + 3.538\ \{\text{RACHS} = 5/6\} + 0.02628\ \text{Year} - 0.0005129(\text{Year-1983})_+^3 + 0.0003196(\text{Year-1988})_+^3 + 0.0004602(\text{Year-1992})_+^3 + 6.697 \times 10^{-5}(\text{Year-1997})_+^3 - 0.0003338(\text{Year-2003})_+^3 - 0.2219\ \text{Age} + 0.001799(\text{Age-0.13})_+^3 - 0.002338(\text{Age-3.05})_+^3 + 0.0005317(\text{Age-12.33})_+^3 + 7.442 \times 10^{-6}(\text{Age-47.61})_+^3 - 1.054 \times 10^{-7}(\text{Age-165})_+^3$ and $\{c\} = 1$ if subject is in group c, 0 otherwise; $(x)_+ = x$ if $x > 0$, 0 otherwise.

The model coefficients and predictive value statistics are shown in Table A2.

TABLE A1. Logistic models of in-hospital death using ABC score as a predictor

					ANOVA	
Model	**Predictor**	**DF**	**Coefficient**	**SE**	***P* (overall)**	***P* (NL)**
Model 1:	ABC score*	4	0.59	0.13	<.0001	.023
C = 0.698			−1.05	0.46		
LR = 490			12.80	6.67		
			−12.98	7.24		
Model 2:	ABC score*	4	0.53	0.15	<.0001	.015
C = 0.737			−0.70	0.52		
LR = 677			7.27	7.52		
			−6.88	8.14		
	Year†	4	−0.03	0.04	<.0001	<.0001
			0.17	0.27		
			−1.11	0.99		
			1.52	1.23		
Model 3:	ABC score*	4	−0.2	0.15	<.0001	<.0001
C = 0.802			2.52	0.58		
LR = 1043			−38.48	8.70		
			41.48	9.46		
	Year†	4	−0.02	0.04	<.0001	<.0001
			−0.13	0.26		
			0.02	0.95		
			0.19	1.23		
	Age (mo)‡	4	−0.31	0.03	<.0001	<.0001
			69.27	12.11		
			−90.32	16.34		
			20.85	4.37		

C, C-index; *LR*, likelihood ratio; *DF*, degrees of freedom; *SE*, standard error; *ANOVA*, analysis of variance; *NL*, nonlinear. *ABC score modeled with a restricted cubic spline (RCS), knot locations: 3, 6, 6.5, 9, and 10.3; †Year of operation modeled with an RCS, knot locations: 1983, 1988, 1992, 1997, 2003; ‡Age in months modeled with an RCS, knot locations: 0.13, 3.05, 12.33, 47.61, and 164.96.

TABLE A2. Logistic models of in-hospital death using RACHS-1 as a predictor

						ANOVA	
Model	**Predictor**	**Level**	**DF**	**Coefficient**	**SE**	***P* overall**	***P* NL**
Model 4	RACHS-1	2	4	1.41	0.24	<.0001	
C = 0.733		3		2.13	0.25		
LR = 667		4		3.18	0.25		
		5/6		4.16	0.26		
Model 5	RACHS-1	2	4	1.49	0.26	<.0001	
C = 0.763		3		2.17	0.25		
LR = 828		4		3.25	0.26		
		5/6		4.52	0.27		
	Year*		4	0.01	0.04	<.0001	<.0001
				0.05	0.26		
				−0.80	0.92		
				1.22	1.18		
Model 6	RACHS-1	2	4	1.37	0.27	<.0001	
C = 0.803		3		1.90	0.26		
LR = 1055		4		2.54	0.26		
		5/6		3.54	0.29		
	Year*		4	0.03	0.04	<.0001	<.0001
				−0.21	0.26		
				0.12	0.95		
				0.18	1.23		
	Age (mo)†		4	−0.22	0.03	<.0001	<.0001
				48.86	11.49		
				−63.50	15.49		
				14.44	4.13		

C, C-index; *LR*, likelihood ratio; *DF*, degrees of freedom; *SE*, standard error; *ANOVA*, analysis of variance; *NL*, nonlinear. *Year of operation modeled with an RCS, knot locations: 1983, 1988, 1992, 1997, 2003; †Age in months modeled with an RCS, knot locations: 0.13, 3.05, 12.33, 47.61, and 164.96.

Discussion

Dr Marc R. de Leval *(London, UK).* I would like to congratulate Dr Al-radi and his colleagues for an important contribution to outcome analysis. The work is a validation study of 2 procedure-adjusted risk stratification methods based both on subjective opinions of a panel of experts. The hospital mortality predicted by the 2 scoring systems is compared with the observed hospital mortality following 13,675 operations performed in a single institution over a 22-year period. Two main findings can be extracted from their analysis.

First, the RACHS-1 categories more consistently represented the probability of hospital deaths compared with the ABC scoring system. We made similar observations in our institution. We assigned the ABC score and the RACHS-1 risk categories to 1085 open cardiac operations performed in the current era. Multiple logistic regression identified RACHS-1 category to be a powerful predictor of mortality, with a P value of $<.0001$, whereas the ABC score was only weakly associated with mortality, with a P value of .03.

The second finding is that both methods are weak discrimination tools in predicting hospital mortality. The authors claim that it is difficult to expect that knowing little else than the procedure, one can accurately predict the outcome. They imply that much more data, both patient- and anomaly specific, would be required. It will be interesting, of course, to see whether the comprehensive ABC score will be a more effective predictor of outcome. We must accept, however, that it will always be impossible to completely predict outcome, and the question is, how complicated should a risk adjustment be?

If the purpose is to be able to compare institutions or individual surgeons, it is important that patient- and procedure-specific factors do not overwhelm potential institution- or surgeon-specific factors. It would be better to try to understand the reasons for variability between institutions that are not going to be explained by minutely detailed case mix adjustment.

I have 2 questions. The first is why do you think that RACHS-1 is superior to the ABC score system in predicting hospital mortality? Do you think that the concept of complexity, which includes technical difficulty, weakens the power of predicting hospital mortality? Today, many technically challenging procedures, such as an arterial switch operation, carry a very small risk of mortality indeed.

And my second question is have you considered putting the 2 scoring systems together in the same equation to find out whether the combination could increase the power of prediction?

Again, I would like to congratulate you for this study and I thank the Association for inviting me to discuss this work.

Dr Osmon O. Al-Radi *(Toronto, Canada).* Dr de Leval, thank you very much for your remarks. Regarding the first question, why RACHS-1 is superior, I think the main advantage of RACHS-1 is that the difference between the highest- and the lowest-risk categories is larger than what it is in ABC. A difference in ABC is about 15% between the lowest- and the highest-risk categories, and the spread between the extreme categories is wider in RACHS-1. The other potential cause is that RACHS-1 in some cases incorporates additional factors other than the operation itself. For example, age in coarctation of the aorta is assigned to a higher-risk category if the patient is older. That is not the case of ABC.

Obviously a more comprehensive score such as the Aristotle comprehensive score will add to the discrimination ability of any tool; however, there is a trade-off between simplicity of use and how much data you need to use the score and whether it would be applicable to data that you have already collected and between how powerful the tool is going to be. You have to establish a balance between how complex you want the score and how powerful do you want it to be. So you have to choose a point that satisfies both the discriminating power and simplicity of use.

In regards to your second question, if you put RACHS-1 and ABC in the same model, RACHS-1 comes out as more predictive. It accounts for all what ABC is telling you. So basically ABC would not be significant if you put them in the same model.

Dr Francois Lacour-Gayet *(Denver, Colo.).* Dr Al-Radi, I have listened with great interest to your presentation. The basic score is the first level of the complexity. It is only a procedure-adjusted complexity, as is RACHS-1. We all know that there are simple Norwood and complex Norwood, simple switch and complex switch. A comprehensive and exhaustive analysis is needed to study individual outcomes.

I will not discuss from a statistical perspective, but intuitively it seems problematic that you ignore in your calculation that there are 4 times the number of patients that could not be analyzed with RACHS-1 compared with ABC.

Finally, constructing a case mix in congenital heart surgery is very challenging. It needs time and attention to detail. We understand that a performance evaluation based on subjective probability and surgical-based knowledge requires a cautious validation. It is in progress. However, today, in absence of validated data in our specialty, if we wait for the data to speak by themselves, there will be only a galactic silence.

Dr Al-Radi. In regards to your first question, there is, again, a balance between how much coverage you want from the scores, whether it covers your entire patient population, and predictive power. You have to establish a balance, again, because if you include patients that have secondary operations, resternotomies, VAD support, that will reduce the predictive power of your score. So, again, it is a balance between how powerful you want the tool to be and the extent of coverage in terms of the procedures that the risk score covers. In regards to your second remark, I have no comment.

Dr Jeffrey H. Silber *(Philadelphia, Pa.).* I am not a cardiac surgeon but I direct the Center for Outcomes Research at The Children's Hospital of Philadelphia and teach severity adjustment at The Wharton School of The University of Pennsylvania, and I really see 2 major problems with this study.

The first is that you used fewer variables to describe the ABC score than you did to describe the RACHS-1 score, and it is very elementary to realize that if you have more variables in a model, you will do a better job fitting the data. Why didn't you fit the ABC score with the same number of variables that you used for the RACHS-1 score? By using fewer variables, you have handicapped the ABC system in your comparisons. The second fundamental problem I see is that you have used different patients to make your comparisons of c-statistics. One of the absolutely essential requirements for comparing severity scores is to use the same patients. By not using the same patients, we really gain very little information as to the comparison between the 2 methods, especially as a larger group of patients were used in the ABC score than the RACHS-1

score. So, not only did you handicap the comparison through your choice of variables, but you also made the comparison meaningless by reporting c-statistics on different populations. I would like to hear your comments on that.

Dr Al-Radi. We did compare RACHS-1 and ABC both as a continuous score and as levels, and we chose the levels for the presentation for the simplicity of the graphs. If you used the continuous score, you would have to use 3-dimensional plots, which I have an example of. The predictive power of ABC did not change whether you used the whole score as a continuous variable or whether you used the ABC as a categorical 4-level variable.

As to your second comment, we also did a sensitivity analysis, including only patients who matched for both scores, and if you do that, the discrimination of the ABC score is somewhat higher but it is still inferior to the RACHS-1.

Dr Silber. Was there a statistical difference between the 2?

Dr Al-Radi. Yes, there was still a statistical significance. But the major point of this presentation is not the comparison between ABC and RACHS-1. I wanted to portray that both scores are short of what would be acceptable as a good method of risk adjustment, and in isolation neither would be adequate for comparing surgeons and institutions. Whether you use RACHS-1 or ABC, you still have to understand that neither is a method that is adequate for complete risk adjustment.

Dr Christo I. Tchervenkov *(Montreal, Canada).* I would just like to raise the issue of the meaning of validation. Simply, the ABC score was based on the opinion of 50 surgeons from across the world, and because the basic premise of the ABC score is that each patient has a constant complexity no matter where in the world this patient is operated, to what extent do you think that the study using data from a single institution has any meaningful significance as to the question of validation?

If you apply the data from another institution that might have a different performance level, then the conclusions may be completely different. What are your comments or thoughts about that and what is it going to take to validate these scores? It perhaps is going to take the data from multiple institutions across different performance levels, different parts of the world.

Thank you very much.

Dr Al-Radi. Our study only addresses 1 aspect of score validity, which is termed *criterion validity* or comparing a score to actual data, and obviously because our data were from a single institution, I do not have the ability to generalize it to a multi-institutional database. If a multi-institutional database was available with the outcomes of interest, then it would be very reasonable to reproduce this work with multi-institutional database. So that would be a very good project.