An empirically based tool for analyzing mortality associated with congenital heart surgery

Sean M. O'Brien, PhD,[a] David R. Clarke, MD,[b] Jeffrey P. Jacobs, MD,[c] Marshall L. Jacobs, MD,[d] Francois G. Lacour-Gayet, MD,[b] Christian Pizarro, MD,[e] Karl F. Welke, MD,[f] Bohdan Maruszewski, MD,[g] Zdzislaw Tobota, MD,[h] Weldon J. Miller, MD,[i] Leslie Hamilton, MD,[j] Eric D. Peterson, MD, MPH,[a] Constantine Mavroudis, MD,[d] and Fred H. Edwards, MD[k]

Objective: Analysis of congenital heart surgery results requires a reliable method of estimating the risk of adverse outcomes. Two major systems in current use are based on projections of risk or complexity that were predominantly subjectively derived. Our goal was to create an objective, empirically based index that can be used to identify the statistically estimated risk of in-hospital mortality by procedure and to group procedures into risk categories.

Methods: Mortality risk was estimated for 148 types of operative procedures using data from 77,294 operations entered into the European Association for Cardiothoracic Surgery (EACTS) Congenital Heart Surgery Database (33,360 operations) and the Society of Thoracic Surgeons (STS) Congenital Heart Surgery Database (43,934 patients) between 2002 and 2007. Procedure-specific mortality rate estimates were calculated using a Bayesian model that adjusted for small denominators. Each procedure was assigned a numeric score (the STS–EACTS Congenital Heart Surgery Mortality Score [2009]) ranging from 0.1 to 5.0 based on the estimated mortality rate. Procedures were also sorted by increasing risk and grouped into 5 categories (the STS–EACTS Congenital Heart Surgery Mortality Categories [2009]) that were chosen to be optimal with respect to minimizing within-category variation and maximizing between-category variation. Model performance was subsequently assessed in an independent validation sample (n = 27,700) and compared with 2 existing methods: Risk Adjustment for Congenital Heart Surgery (RACHS-1) categories and Aristotle Basis Complexity scores.

Results: Estimated mortality rates ranged across procedure types from 0.3% (atrial septal defect repair with patch) to 29.8% (truncus plus interrupted aortic arch repair). The proposed STS–EACTS score and STS–EACTS categories demonstrated good discrimination for predicting mortality in the validation sample (C-index = 0.784 and 0.773, respectively). For procedures with more than 40 occurrences, the Pearson correlation coefficient between a procedure's STS–EACTS score and its actual mortality rate in the validation sample was 0.80. In the subset of procedures for which RACHS-1 and Aristotle Basic Complexity scores are defined, discrimination was highest for the STS–EACTS score (C-index = 0.787), followed by STS–EACTS categories (C-index = 0.778), RACHS-1 categories (C-index = 0.745), and Aristotle Basic Complexity scores (C-index = 0.687). When patient covariates were added to each model, the C-index improved: STS–EACTS score (C-index = 0.816), STS–EACTS categories (C-index = 0.812), RACHS-1 categories (C-index = 0.802), and Aristotle Basic Complexity scores (C-index = 0.795).

Conclusion: The proposed risk scores and categories have a high degree of discrimination for predicting mortality and represent an improvement over existing consensus-based methods. Risk models incorporating these measures may be used to compare mortality outcomes across institutions with differing case mixes.

Cardiac surgeons have recognized and emphasized the need to establish clinical registries and quantitative tools for responsible reporting of outcomes. Large multi-institutional databases, such as the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database, among others, have developed, applied, and validated methods of risk adjustment in reporting outcomes. This has addressed appropriate concerns that the reporting of raw, unadjusted mortality data is misleading and potentially penalizes surgeons and centers

From the Duke Clinical Research Institute,[a] Durham, NC; the Children's Hospital Heart Institute,[b] Denver, Colo; the Congenital Heart Institute of Florida (CHIF),[c] Saint Petersburg and Tampa, Fla; The Cleveland Clinic,[d] Cleveland, Ohio; the Nemours Cardiac Center,[e] Alfred I. duPont Hospital for Children, Wilmington, Del; the Oregon Health and Science University,[f] Portland, Ore; Memorial Hospital Child's Health Centre,[g] Warsaw, Poland; Children's Memorial Health Institute,[h] Warsaw, Poland; Rho, Inc,[i] Chapel Hill, NC; Freeman Hospital,[j] Newcastle upon Tyne, United Kingdom; and the University of Florida,[k] Jacksonville, Fla.

Read at the Thirty-fourth Annual Meeting of The Western Thoracic Surgical Association, Kona, Hawaii, June 25–28, 2008.

Received for publication June 20, 2008; revisions received Nov 18, 2008; accepted for publication March 7, 2009.

Address for reprints: Sean M. O'Brien, PhD, Box 17969, Duke Clinical Research Institute, Durham, NC 27715 (E-mail: obrie027@mc.duke.edu).

J Thorac Cardiovasc Surg 2009;138:1139-53
0022-5223/$36.00

doi:10.1016/j.jtcvs.2009.03.071

Abbreviations and Acronyms

ABC = Aristotle Basic Complexity
EACTS = European Association for Cardiothoracic Surgery
RACHS-1 = Risk Adjustment for Congenital Heart Surgery
STS = Society of Thoracic Surgeons

that manage high-risk patients and complex procedures because observed mortality rates might be higher than in centers dealing with less challenging cases. The kinds of statistical tools and risk models that have been developed to address these issues when the clinical substrate is adult patients with acquired cardiovascular disease cannot simply be applied to the population of pediatric and adult patients with congenital heart disease. Here the problem is considerably more complex, in large part because the individual diagnoses and distinct types of surgical procedures number in the hundreds, despite the fact that the universe of patients with congenital heart disease is considerably smaller than that of adult patients with ischemic and valvular heart disease. As a result, the number of patients in some diagnostic and procedural groups is quite small. Nonetheless, it is recognized that the need to establish tools for case-mix adjustment is fundamental to any systematic attempt to measure outcomes, compare performance, and sustain a program of continual quality improvement.

As a response to the need for case-mix adjustment of outcome data but in the absence of significant amounts of registry data in 2000, the Aristotle Complexity score was developed.[1,2] Using the expert opinions of 50 internationally based surgeons, the Aristotle Basic Complexity (ABC) score was constructed for 145 distinct congenital heart surgery procedures. Three components (potential for mortality, potential for morbidity, and technical difficulty) were subjectively scored, and the sum became the ABC score.

Separately, another group of researchers developed the Risk Adjustment for Congenital Heart Surgery (RACHS-1) system, also using an expert panel.[3,4] RACHS-1 groups procedures into 6 levels of increasing risk of mortality. This allocation of procedures was subsequently refined using empirical data from 2 multi-institutional registries. When compared with the ABC score, the RACHS-1 categories appear to have better discrimination for predicting mortality, whereas the ABC score covers a larger proportion of congenital heart surgery case volume.[5-7]

The largest validation study of the ABC score was recently conducted by using a combined sample of nearly 36,000 patients from the STS Congenital Heart Surgery Database and the European Association for Cardiothoracic Surgery (EACTS) Congenital Heart Surgery Database.[7] In that study there was a significant increasing association between the ABC score and in-hospital mortality, with an overall C-index of 0.70. Although it was clear that the ABC score generally discriminated between low-risk and high-risk procedures, it was also clear that for a relatively small number of individual procedures, the initial estimation of mortality risk by the Aristotle international panel of surgical experts did not accurately predict the actual empirical estimates observed over the ensuing decade.

The goal of the present study was to derive a new system for classifying congenital heart surgery procedures based on their potential for in-hospital mortality using empirical data from the STS and EACTS databases. There were 3 specific objectives.

First, we sought to estimate procedure-specific relative risks of in-hospital mortality using a statistical model that accounts for uncertainty in procedures with small sample sizes.

Second, we sought to convert these procedure-specific mortality estimates into a scale ranging from 0.1 to 5.0. The range of this scale was chosen for consistency with the Aristotle method. The resulting score has been named the STS–EACTS Congenital Heart Surgery Mortality Score (2009) (or, briefly, the STS–EACTS score).

Third, we sought to group procedures with similar estimated mortality risk into a small number of relatively homogeneous categories (the STS–EACTS Congenital Heart Surgery Mortality Categories [2009] or, briefly, the STS–EACTS categories). These categories are intended to serve as a stratification variable that can be used to adjust for case mix when analyzing outcomes and comparing institutions.

MATERIALS AND METHODS

Study Population

The STS Congenital Heart Surgery Database and the EACTS Database are described elsewhere.[8] The study population consisted of patients who underwent a congenital cardiovascular operation at an STS-participating hospital between January 1, 2002, and December 31, 2006, or at an EACTS-participating hospital between January 1, 2002, and April 4, 2007. Data from 1 STS center were excluded because this participant did not consistently report outcomes during the study period. Only the first operation of each hospital admission was analyzed. Operations were included if they involved one of the 148 cardiovascular procedures listed in Table 1. This list includes all cardiovascular procedures that were included in the short-list nomenclature of the STS and EACTS databases and appeared at least once as the primary procedure of an operation in the STS–EACTS dataset. Patients weighing less than or equal to 2500 g undergoing patent ductus arteriosus ligation as their primary procedure were excluded from the analysis because they are not included in mortality calculations in the EACTS and STS Congenital Database reports. In addition, 244 (0.3%) patients with missing in-hospital mortality status were excluded. The final study population consisted of 43,934 operations from 57 centers in the STS database and 33,360 operations from 91 centers in the EACTS database for a total of 77,294 operations.

The risk tool developed using this dataset was subsequently validated in a separate sample of STS and EACTS patients meeting the same inclusion criteria described above. This validation sample consisted of 20,042 operations performed between January 1, 2007, and June 30, 2008, in the STS database and 7658 operations performed between April 5, 2007, and April 8, 2008, in the EACTS database.

TABLE 1. Procedure names, proposed scores and categories, and data for model development

	Procedure scores			No. of operations		Estimated mortality risk	
Procedure name	Difficulty ranking	Mortality score	Mortality category	All operations	No. with nonmissing mortality	Unadjusted % (95% interval*)	Model based % (95% interval†)
ASD repair, patch	8	0.1	1	4035	4028	0.2% (0.1%–0.4%)	0.3% (0.1%–0.5%)
AVC (AVSD) repair, partial (incomplete) (PAVSD)	31	0.1	1	1064	1062	0.3% (0.1%–0.8%)	0.5% (0.2%–0.9%)
ASD repair, patch + PAPCV repair	28	0.2	1	438	438	0.2% (0.0%–1.3%)	0.6% (0.2%–1.4%)
Aortic stenosis, subvalvar, repair	42	0.2	1	1834	1828	0.5% (0.3%–1.0%)	0.6% (0.3%– 1.0%)
ICD (AICD) implantation	14	0.2	1	391	384	0.3% (0.0%–1.4%)	0.7% (0.2%–1.6%)
DCRV repair	48	0.2	1	467	467	0.4% (0.1%–1.5%)	0.8% (0.2%–1.6%)
ASD repair, primary closure	7	0.2	1	2230	2229	0.8% (0.5%–1.3%)	0.9% (0.5%–1.3%)
VSD repair, patch	32	0.2	1	6717	6702	0.9% (0.7%–1.1%)	0.9% (0.7%–1.1%)
Vascular ring repair	19	0.2	1	899	895	0.8% (0.3%–1.6%)	0.9% (0.4%–1.6%)
Coarctation repair, end to end	24	0.2	1	1703	1702	0.9% (0.5%– 1.5%)	1.0% (0.6%–1.5%)
ICD (AICD) procedure	15	0.2	1	127	126	0.0% (0.0%–2.9%)	1.0% (0.2%–2.9%)
PFO, primary closure	6	0.2	1	217	216	0.5% (0.0%–2.6%)	1.1% (0.3%–2.5%)
AVR, bioprosthetic	55	0.3	1	101	101	0.0% (0.0%–3.6%)	1.2% (0.2%–3.4%)
VSD repair, primary closure	30	0.3	1	754	752	1.1% (0.5%–2.1%)	1.2% (0.6%–2.1%)
PVR	44	0.3	1	682	680	1.2% (0.5%–2.3%)	1.3% (0.6%–2.3%)
Conduit reoperation	77	0.3	1	1303	1299	1.3% (0.8%–2.1%)	1.4% (0.8%–2.1%)
Pacemaker procedure	3	0.3	1	1411	1408	1.3% (0.8%–2.1%)	1.4% (0.9%–2.1%)
PAPVC repair	27	0.3	1	481	481	1.2% (0.5%–2.7%)	1.5% (0.7%–2.7%)
TOF repair, ventriculotomy, nontransanular patch	62	0.3	1	930	928	1.4% (0.7%–2.4%)	1.5% (0.8%–2.4%)
TOF repair, no ventriculotomy	81	0.3	1	862	860	1.4% (0.7%–2.4%)	1.5% (0.8%–2.3%)
Glenn (unidirectional cavopulmonary anastomosis; unidirectional Glenn procedure)	41	0.3	1	65	65	0.0% (0.0%–5.5%)	1.5% (0.2%–4.3%)
AVC (AVSD) repair, intermediate (transitional)	33	0.3	1	421	420	1.4% (0.5%–3.1%)	1.6% (0.7%–3.0%)
Coarctation repair, interposition graft	49	0.3	1	114	114	0.9% (0.0%–4.8%)	1.7% (0.4%–4.1%)
Fontan, TCPC, lateral tunnel, fenestrated	101	0.3	1	743	742	1.6% (0.8%–2.8%)	1.7% (0.9%–2.7%)
Sinus of Valsalva, aneurysm repair	61	0.3	1	53	53	0.0% (0.0%–6.7%)	1.7% (0.3%–5.2%)
AVR, mechanical	52	0.3	1	384	383	1.6% (0.6%–3.4%)	1.7% (0.7%–3.2%)
PDA closure, surgical	5	0.4	2	1922	1910	1.8% (1.3%–2.5%)	1.9% (1.3%–2.5%)
PA, reconstruction (plasty), main (trunk)	25	0.4	2	192	191	1.6% (0.3%–4.5%)	1.9% (0.6%–4.0%)
LV to aorta tunnel repair	90	0.4	2	42	42	0.0% (0.0%–8.4%)	1.9% (0.3%–5.9%)
Valvuloplasty, mitral	76	0.4	2	1751	1747	1.9% (1.3%–2.6%)	1.9% (1.3%–2.6%)
Valvuloplasty, aortic	72	0.4	2	861	861	1.9% (1.1%–3.0%)	1.9% (1.1%–2.9%)
11/2 Ventricular repair	58	0.4	2	39	39	0.0% (0.0%–9.0%)	2.0% (0.3%–6.2%)

TABLE 1. Continued

Procedure name	Procedure scores: Difficulty ranking	Procedure scores: Mortality score	Procedure scores: Mortality category	No. of operations: All operations	No. of operations: No. with nonmissing mortality	Estimated mortality risk: Unadjusted % (95% interval*)	Estimated mortality risk: Model based % (95% interval†)
Arrhythmia surgery–ventricular, surgical ablation	85	0.4	2	33	33	0.0% (0.0%–10.6%)	2.2% (0.3%–6.8%)
Pacemaker implantation, permanent	2	0.4	2	1086	1077	2.1% (1.4%–3.2%)	2.2% (1.4%–3.1%)
Ross procedure	127	0.4	2	620	617	2.1% (1.1%–3.6%)	2.2% (1.3%–3.4%)
Glenn+PA reconstruction	71	0.4	2	428	426	2.1% (1.0%–4.0%)	2.2% (1.1%–3.8%)
Aortopexy	4	0.4	2	30	30	0.0% (0.0%–11.6%)	2.3% (0.3%–7.3%)
Fontan, atriopulmonary connection	94	0.4	2	30	30	0.0% (0.0%–11.6%)	2.3% (0.3%–6.9%)
Bilateral bidirectional cavopulmonary anastomosis (bilateral bidirectional Glenn procedure)	63	0.4	2	449	449	2.2% (1.1%–4.1%)	2.4% (1.2%–3.8%)
Aortic root replacement, mechanical	111	0.5	2	145	145	2.1% (0.4%–5.9%)	2.4% (0.7%–5.1%)
Conduit placement, LV to PA	73	0.5	2	25	25	0.0% (0.0%–13.7%)	2.4% (0.3%–7.9%)
Coarctation repair, end to end, extended	50	0.5	2	1965	1961	2.5% (1.9%–3.3%)	2.5% (1.9%–3.3%)
Anomalous origin of coronary artery repair	119	0.5	2	327	326	2.5% (1.1%–4.8%)	2.6% (1.2%–4.4%)
RVOT procedure	40	0.5	2	1591	1583	2.6% (1.9%–3.5%)	2.6% (1.9%–3.5%)
Aortic aneurysm repair	93	0.5	2	322	321	2.5% (1.1%–4.9%)	2.6% (1.3%–4.5%)
Congenitally corrected TGA repair, VSD closure	106	0.5	2	21	21	0.0% (0.0%–16.1%)	2.6% (0.3%–8.8%)
AP window repair	35	0.5	2	125	125	2.4% (0.5%–6.9%)	2.7% (0.9%–5.6%)
Valvuloplasty, pulmonic	26	0.5	2	307	307	2.6% (1.1%–5.1%)	2.7% (1.3%–4.7%)
TOF repair, ventriculotomy, transannular patch	79	0.5	2	2541	2535	2.7% (2.1%–3.4%)	2.7% (2.1%–3.4%)
Aortic root replacement, bioprosthetic	120	0.5	2	20	20	0.0% (0.0%–16.8%)	2.7% (0.3%–9.3%)
Bidirectional cavopulmonary anastomosis (bidirectional Glenn procedure)	43	0.5	2	2502	2492	2.7% (2.1%–3.4%)	2.7% (2.1%–3.4%)
Aortic stenosis, supravalvar, repair	64	0.5	2	336	335	2.7% (1.2%–5.0%)	2.8% (1.4%–4.6%)
Pericardiectomy	20	0.5	2	48	48	2.1% (0.1%–11.1%)	2.9% (0.5%–7.5%)
Conduit placement, other	75	0.5	2	16	16	0.0% (0.0%–20.6%)	2.9% (0.3%–9.8%)
Aneurysm, ventricular, left, repair	107	0.5	2	47	46	2.2% (0.1%–11.5%)	3.0% (0.5%–7.8%)
Fontan, TCPC, external conduit, fenestrated	96	0.6	2	1241	1238	3.0% (2.1%–4.1%)	3.0% (2.1%–4.0%)
Pulmonary artery origin from ascending aorta (hemitruncus) repair	89	0.6	2	43	43	2.3% (0.1%–12.3%)	3.1% (0.6%–8.2%)

TABLE 1. Continued

Procedure name	Procedure scores			No. of operations		Estimated mortality risk	
	Difficulty ranking	Mortality score	Mortality category	All operations	No. with nonmissing mortality	Unadjusted % (95% interval*)	Model based % (95% interval†)
ASD, common atrium (single atrium), septation	18	0.6	2	44	44	2.3% (0.1%–12.0%)	3.1% (0.5%–8.3%)
PAPVC, scimitar, repair	91	0.6	2	72	72	2.8% (0.3%–9.7%)	3.2% (0.8%–7.7%)
Fontan, TCPC, external conduit, nonfenestrated	97	0.6	2	809	807	3.2% (2.1%–4.7%)	3.2% (2.1%–4.6%)
Ligation, pulmonary artery	16	0.6	2	11	11	0.0% (0.0%–28.5%)	3.4% (0.4%–12.1%)
Coronary artery fistula ligation	17	0.6	2	39	38	2.6% (0.1%–13.8%)	3.4% (0.6%–9.2%)
Aortic root replacement, valve sparing	142	0.6	2	37	37	2.7% (0.1%–14.2%)	3.4% (0.6%–9.2%)
Mitral stenosis, supravalvar mitral ring repair	74	0.6	2	86	86	3.5% (0.7%–9.9%)	3.6% (1.0%–7.7%)
Arrhythmia surgery–atrial, surgical ablation	84	0.7	2	273	272	3.7% (1.8%–6.7%)	3.6% (1.9%–5.9%)
Systemic venous stenosis repair	56	0.7	2	59	59	3.4% (0.4%–11.7%)	3.7% (0.9%–8.6%)
PA, reconstruction (plasty), branch, peripheral (at or beyond the hilar bifurcation)	70	0.7	2	189	189	3.7% (1.5%–7.5%)	3.7% (1.6%–6.5%)
Valvuloplasty, tricuspid	57	0.7	2	1182	1178	3.7% (2.7%–5.0%)	3.7% (2.8%–4.9%)
TVR	65	0.7	2	133	133	3.8% (1.2%–8.6%)	3.8% (1.5%–7.3%)
Valve replacement, truncal valve	46	0.7	2	8	8	0.0% (0.0%–36.9%)	3.8% (0.4%–13.8%)
Fontan, TCPC, lateral tunnel, nonfenestrated	99	0.7	2	104	104	3.8% (1.1%–9.6%)	3.9% (1.3%–7.9%)
Atrial fenestration closure	38	0.7	2	29	29	3.4% (0.1%–17.8%)	3.9% (0.7%–11.3%)
Cor triatriatum repair	60	0.7	2	177	176	4.0% (1.6%–8.0%)	4.0% (1.8%–7.2%)
VSD, multiple, repair	113	0.7	2	325	324	4.0% (2.2%–6.8%)	4.0% (2.2%–6.3%)
Atrial baffle procedure (non-Mustard, non-Senning)	67	0.7	2	26	26	3.8% (0.1%–19.6%)	4.0% (0.7%–11.0%)
Coarctation repair, subclavian flap	23	0.7	2	219	219	4.1% (1.9%–7.7%)	4.1% (2.0%–6.9%)
Partial left ventriculectomy (LV volume reduction surgery; Batista)	133	0.7	2	26	26	3.8% (0.1%–19.6%)	4.1% (0.7%–11.3%)
TOF repair, RV–PA conduit	80	0.7	2	362	358	4.2% (2.4%–6.8%)	4.2% (2.4%–6.4%)
Transplantation, lung(s)	129	0.8	3	94	93	4.3% (1.2%–10.6%)	4.2% (1.4%–8.6%)
Occlusion MAPCA(s)	51	0.8	3	26	26	3.8% (0.1%–19.6%)	4.2% (0.7%–12.1%)
Coarctation repair + VSD repair	112	0.8	3	329	327	4.3% (2.4%–7.1%)	4.2% (2.4%–6.6%)
Konno procedure	131	0.8	3	162	162	4.3% (1.8%–8.7%)	4.3% (1.9%–7.6%)
Coarctation repair, patch aortoplasty	22	0.8	3	395	393	4.3% (2.5%–6.8%)	4.3% (2.6%–6.5%)

TABLE 1. Continued

Procedure name	Procedure scores			No. of operations		Estimated mortality risk	
	Difficulty ranking	Mortality score	Mortality category	All operations	No. with nonmissing mortality	Unadjusted % (95% interval*)	Model based % (95% interval†)
PA, reconstruction (plasty), branch, central (within the hilar bifurcation)	68	0.8	3	646	644	4.3% (2.9%–6.2%)	4.3% (2.9%–5.9%)
Aneurysm, pulmonary artery, repair	53	0.8	3	23	23	4.3% (0.1%–21.9%)	4.3% (0.8%–12.2%)
Aneurysm, ventricular, right, repair	86	0.8	3	91	91	4.4% (1.2%–10.9%)	4.3% (1.4%–8.8%)
Ventricular septal fenestration	45	0.8	3	24	24	4.2% (0.1%–21.1%)	4.4% (0.8%–12.4%)
Shunt, ligation and takedown	11	0.8	3	65	65	4.6% (1.0%–12.9%)	4.5% (1.3%–9.9%)
Hemi-Fontan procedure	78	0.8	3	262	260	4.6% (2.4%–7.9%)	4.5% (2.4%–7.1%)
AVC (AVSD) repair, complete	87	0.8	3	2869	2860	4.6% (3.9%–5.4%)	4.6% (3.9%–5.4%)
Anomalous systemic venous connection repair	54	0.8	3	166	166	4.8% (2.1%–9.3%)	4.8% (2.2%–8.2%)
ASO	115	0.8	3	2069	2068	4.8% (3.9%–5.8%)	4.8% (3.9%–5.7%)
Valvuloplasty, truncal valve	59	0.8	3	20	20	5.0% (0.1%–24.9%)	4.8% (0.8%–13.5%)
Fontan, atrioventricular connection	102	0.9	3	2	2	0.0% (0.0%–84.2%)	4.9% (0.4%–20.1%)
Pulmonary embolectomy, acute pulmonary embolus	34	0.9	3	2	2	0.0% (0.0%–84.2%)	5.0% (0.4%–19.7%)
ASD partial closure	10	0.9	3	37	37	5.4% (0.7%–18.2%)	5.1% (1.1%–12.7%)
Rastelli operation	125	0.9	3	333	333	5.4% (3.2%–8.4%)	5.3% (3.2%–7.8%)
Conduit placement, ventricle to aorta	95	0.9	3	1	1	0.0% (0.0%–97.5%)	5.3% (0.5%–21.4%)
AVR, homograft	110	1	3	30	30	6.7% (0.8%–22.1%)	5.8% (1.3%–13.8%)
REV	126	1.1	3	26	26	7.7% (0.9%–25.1%)	6.3% (1.3%–15.5%)
Pulmonary artery sling repair	105	1.1	3	88	86	7.0% (2.6%–14.6%)	6.4% (2.5%–11.9%)
Mustard procedure	100	1.1	3	25	25	8.0% (1.0%–26.0%)	6.4% (1.4%–15.9%)
Pulmonary atresia–VSD (including TOF, PA) repair	92	1.1	3	289	289	6.6% (4.0%–10.1%)	6.4% (4.0%–9.3%)
Conduit placement, RV to PA	66	1.2	3	965	964	6.7% (5.2%–8.5%)	6.7% (5.2%–8.4%)
Pulmonary embolectomy	37	1.2	3	9	9	11.1% (0.3%–48.2%)	7.1% (1.0%–22.1%)
MVR	69	1.3	4	637	636	7.4% (5.5%–9.7%)	7.3% (5.4%–9.4%)
Pericardial drainage procedure	1	1.3	4	258	256	7.8% (4.8%–11.8%)	7.5% (4.7%–11.0%)
Aortic arch repair	82	1.4	4	787	782	7.9% (6.1%–10.0%)	7.8% (6.1%–9.8%)
Fontan revision or conversion (redo Fontan procedure)	143	1.4	4	68	68	8.8% (3.3%–18.2%)	7.9% (3.1%–14.6%)
DOLV repair	130	1.4	4	7	7	14.3% (0.4%–57.9%)	7.9% (1.0%–24.0%)
DORV, intraventricular tunnel repair	132	1.4	4	583	582	8.1% (6.0%–10.6%)	8.0% (6.0%–10.3%)

TABLE 1. Continued

Procedure name	Procedure scores			No. of operations		Estimated mortality risk	
	Difficulty ranking	Mortality score	Mortality category	All operations	No. with nonmissing mortality	Unadjusted % (95% interval*)	Model based % (95% interval†)
Arterial switch procedure + aortic arch repair	136	1.4	4	18	18	11.1% (1.4%–34.7%)	8.0% (1.7%–20.6%)
PA debanding	29	1.4	4	104	104	8.7% (4.0%–15.8%)	8.0% (3.7%–13.7%)
ASO and VSD repair	138	1.4	4	987	985	8.3% (6.7%–10.2%)	8.2% (6.6%–10.0%)
Cardiac tumor resection	88	1.4	4	221	220	8.6% (5.3%–13.2%)	8.3% (5.1%–12.2%)
Transplantation, heart	103	1.4	4	626	625	8.5% (6.4%–10.9%)	8.4% (6.3%–10.6%)
Coronary artery bypass	98	1.5	4	62	62	9.7% (3.6%–19.9%)	8.5% (3.5%–16.0%)
TOF–absent pulmonary valve repair	109	1.5	4	166	165	9.1% (5.2%–14.6%)	8.6% (5.0%–13.1%)
Valve excision, tricuspid (without replacement)	13	1.5	4	5	5	20.0% (0.5%–71.6%)	8.8% (1.2%–28.1%)
Shunt, systemic to pulmonary, MBTS	39	1.5	4	2793	2785	8.9% (7.9%–10.1%)	8.9% (7.9%– 10.0%)
TOF–AVC (AVSD) repair	122	1.6	4	145	144	9.7% (5.4%–15.8%)	9.1% (5.0%–14.1%)
Ross–Konno procedure	146	1.6	4	205	205	9.8% (6.1%–14.7%)	9.4% (5.8%–13.9%)
Senning procedure	108	1.6	4	45	45	11.1% (3.7%–24.1%)	9.4% (3.5%–18.6%)
Ebstein's repair	124	1.6	4	65	65	10.8% (4.4%–20.9%)	9.5% (4.0%–17.6%)
Aortic arch repair + VSD repair	123	1.7	4	339	338	10.1% (7.1%–13.8%)	9.8% (6.9%–13.1%)
PA banding	21	1.7	4	1298	1292	9.9% (8.3%–11.7%)	9.8% (8.3%–11.5%)
Aortic root replacement, homograft	121	1.7	4	104	102	10.8% (5.5%–18.5%)	9.9% (5.1%–16.2%)
Unifocalization MAPCA(s)	116	1.7	4	319	319	10.3% (7.2%–14.2%)	10.0% (7.1%–13.4%)
Aortic dissection repair	128	1.7	4	32	31	12.9% (3.6%–29.8%)	10.0% (3.0%–21.1%)
Congenitally corrected TGA repair, VSD closure and LV to PA conduit	135	1.7	4	12	12	16.7% (2.1%–48.4%)	10.1% (2.0%–25.9%)
Pulmonary atresia–VSD–MAPCA (pseudotruncus) repair	137	1.7	4	160	158	10.8% (6.4%–16.7%)	10.2% (6.1%–15.3%)
VSD creation/enlargement	83	1.8	4	107	106	11.3% (6.0%–18.9%)	10.4% (5.6%–16.6%)
HLHS biventricular repair	145	1.9	4	64	64	12.5% (5.6%–23.2%)	10.9% (4.8%– 18.8%)
TAPVC repair	104	1.9	4	1381	1379	11.2% (9.6%–13.0%)	11.2% (9.5%–12.8%)
Pulmonary venous stenosis repair	117	2	4	270	268	11.9% (8.3%–16.4%)	11.4% (8.0%–15.3%)
Shunt, systemic to pulmonary, central (from aorta or to main pulmonary artery)	47	2.1	4	663	661	12.3% (9.9%–15.0%)	12.1% (9.7%–14.6%)
Interrupted aortic arch repair	118	2.1	4	519	515	12.4% (9.7%–15.6%)	12.2% (9.6%–15.1%)
Arterial switch procedure and VSD repair + aortic arch repair	144	2.4	4	113	113	15.0% (9.0%–23.0%)	14.0% (8.5%–20.5%)
Truncus arteriosus repair	134	2.4	4	592	586	14.3% (11.6%–17.4%)	14.1% (11.4%–16.8%)
ASD creation/enlargement	9	2.5	4	138	136	15.4% (9.8%–22.6%)	14.5% (9.4%–20.9%)
Atrial septal fenestration	12	2.6	4	18	18	22.2% (6.4%–47.6%)	15.1% (4.5%–30.8%)

CHD

TABLE 1. Continued

Procedure name	Procedure scores: Difficulty ranking	Procedure scores: Mortality score	Procedure scores: Mortality category	No. of operations: All operations	No. of operations: No. with nonmissing mortality	Estimated mortality risk: Unadjusted % (95% interval*)	Estimated mortality risk: Model based % (95% interval†)
Valve closure, tricuspid (exclusion, univentricular approach)	36	2.6	4	5	5	40.0% (5.3%–85.3%)	15.6% (2.7%–41.6%)
Damus–Kaye–Stansel procedure (creation of AP anastomosis without arch reconstruction)	114	2.9	5	344	343	17.5% (13.6%–21.9%)	17.1% (13.2%–21.5%)
Transplantation, heart and lung	141	3.2	5	13	13	30.8% (9.1%–61.4%)	18.7% (5.4%–39.8%)
Congenitally corrected TGA repair, atrial switch and Rastelli operation	139	3.2	5	18	18	27.8% (9.7%–53.5%)	18.9% (6.3%–37.2%)
Congenitally corrected TGA repair, atrial switch and ASO (double switch)	148	3.4	5	32	32	25.0% (11.5%–43.4%)	20.0% (9.1%–34.7%)
Norwood procedure	147	4	5	2383	2359	23.7% (22.0%–25.4%)	23.6% (21.9%–25.3%)
Truncus + IAA repair	140	5	5	43	43	34.9% (21.0%–50.9%)	29.8% (17.7%–44.3%)

ASD, Atrial septal defect; *AVC*, atrioventricular canal; *AVSD*, atrioventricular septal defect; *PAVSD*, partial atrioventricular septal defect; *PAPVC*, partial anomalous pulmonary venous connection; *ICD*, implantable cardioverter defibrillator; *AICD*, automatic implantable cardioverter defibrillator; *DCRV*, double-chambered right ventricle; *VSD*, ventricular septal defect; *PFO*, patent foramen ovale; *AVR*, aortic valve replacement; *PVR*, pulmonary valve replacement; *TOF*, tetralogy of Fallot; *TCPC*, total cavopulmonary connection; *PDA*, patent ductus arteriosus; *PA*, pulmonary artery; *LV*, left ventricle; *RVOT*, right ventricular outflow tract; *TGA*, transposition of the great arteries; *AP*, aortopulmonary; *TVR*, tricuspid valve replacement; *RV*, right ventricle; *MAPCA*, major aortopulmonary collateral artery; *ASO*, arterial switch operation; *REV*, réparation à l'étage ventriculaire (REV procedure); *MVR*, mitral valve replacement; *DOLV*, double-outlet left ventricle; *MBTS*, modified Blalock–Taussig shunt; *HLHS*, hypoplastic left heart syndrome; *TAPVC*, total anomalous pulmonary venous connection; *IAA*, interrupted aortic arch. *Denotes 95% exact binomial confidence interval. †Denotes 95% Bayesian credible interval.

Hospitals participating in the STS and EACTS registries are required to comply with local regulatory and privacy guidelines. The Duke Clinical Research Institute serves as the data analysis center for the STS database and has an agreement, as well as institutional review board approval, to analyze the aggregate deidentified data for research purposes.

Classification of Multiple-Procedure Operations

Several procedures listed in Table 1 are actually combinations of 2 or more procedures. These combinations were identified by the Aristotle expert panel because they occur frequently in the STS and EACTS databases and because the complexity of the combination is regarded as being different from the complexity of the component procedures when performed in isolation. For all other operations involving combinations of procedures, the operation was classified according to the most technically complex procedure, as determined by the difficulty component of the 2007 update of the ABC score. The ABC score contains some ties and is not defined for 3 of the procedures listed in Table 1. To deal with undefined or tied Aristotle scores, 6 of the study authors independently ranked the difficulty of each procedure listed in Table 1. Undefined or tied Aristotle scores were adjudicated by assigning the operation to the procedure with the highest average ranking determined by the 6 graders. The difficulty rankings are included in Table 1 so that users of the risk tool will be able to replicate our method of classifying multiple-procedure operations.

End Point

The study end point was in-hospital mortality, which was defined as death during the same hospitalization as surgery regardless of cause.

Estimation of Procedure-Specific Mortality Rates

Mortality estimates were calculated by using a Bayesian random effects model that adjusted each procedure's mortality rate based on the size of the denominator. Using a statistical model was considered advantageous because several individual procedures had small denominators, and hence their unadjusted mortality rates were susceptible to chance fluctuations. Unlike conventional methods, random effects models use data from all of the procedures in the database when estimating the probability of mortality for any single procedure. This ''borrowing of information'' across procedures produces estimates with good statistical properties, including smaller standard errors than conventional estimates. Heuristically, the model-based estimate is a weighted average of a procedure's actual observed mortality rate and the overall average mortality rate for all procedures in the database. The model weights an individual procedure's own data more heavily when the denominator is large enough to be reliable and weights the overall average mortality rate more heavily when the denominator is too small to support a reliable mortality estimate. For procedures with more than 200 occurrences, the model-based estimates were virtually identical to the usual unadjusted (raw) mortality percentages (Appendix 1).

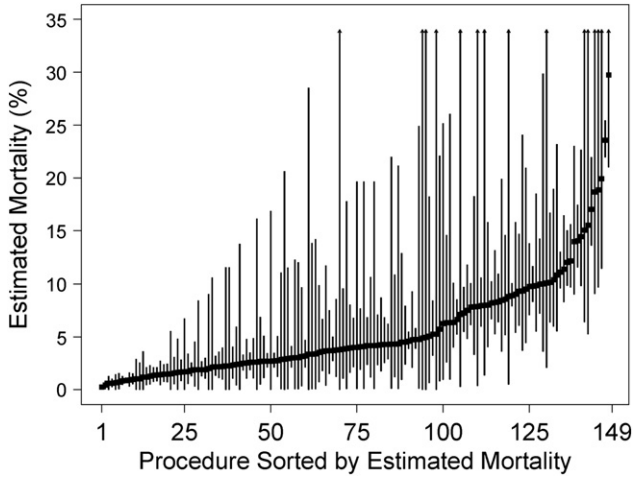

FIGURE 1. Procedure-specific estimated mortality rates. *Square dots* represent model-based procedure-specific mortality estimates. *Vertical lines* represent exact 95% binomial confidence intervals.

Creation of the Mortality Score

Each procedure was assigned a numeric score (STS–EACTS score) ranging from 0.1 to 5.0. The scores were assigned by shifting and rescaling the estimated procedure-specific mortality rates to lie in the interval from 0.1 to 5.0 and then rounding to one decimal place. The following formula was used:

$$\text{Mortality score of } j\text{-th procedure} = 0.1 + 4.9 \times \frac{p_j - \min}{\max - \min},$$

where p_j denotes the estimated risk of the *j-th* procedure, and *max* and *min* denote the maximum and minimum values of p_j across the 148 procedures.

Creation of Mortality Categories

Procedures were sorted by increasing estimated risk and partitioned into 5 relatively homogeneous categories (STS–EACTS categories). Five categories was the smallest number that did not result in excessive within-category heterogeneity. Within-category homogeneity was measured objectively using a weighted sum of squares criterion (Appendix 2).[9] A dynamic programming algorithm was then used to find the categorization that maximizes the homogeneity criterion. This data-driven approach ensures that procedures in the same category will be as similar as possible with respect to their estimated mortality risk.

To determine the number of categories, we evaluated the performance of different categorizations consisting of 2 to 20 categories. Performance was assessed internally based on 2 criteria. First, we evaluated the internal homogeneity of the categories using the criterion described in Appendix 2. Second, we assessed the discrimination of the categories as predictors of mortality. Discrimination was quantified by the area under the receiver operating characteristic curve (also known as the C-index).[10] The C-index is interpreted as the probability that a randomly selected patient who died was considered to be higher risk than a randomly selected patient who survived. The C-index generally ranges from 0.5 to 1.0, with 0.5 representing no discrimination (ie, a coin flip) and 1.0 representing perfect discrimination.

TABLE 2. Characteristics of proposed risk categories in 2002–2007 STS and EACTS data

	STS–EACTS mortality category				
	1	2	3	4	5
Range of scores	0.1–0.3	0.4–0.7	0.8–1.2	1.3–2.6	2.7–5.0
No. of procedures	26	52	27	37	6
No. of patients	28,363	23,235	9026	13,862	2808
No. of deaths	234	601	449	1374	650
Mortality	0.8%	2.6%	5.0%	9.9%	23.1%

STS–EACTS, Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery.

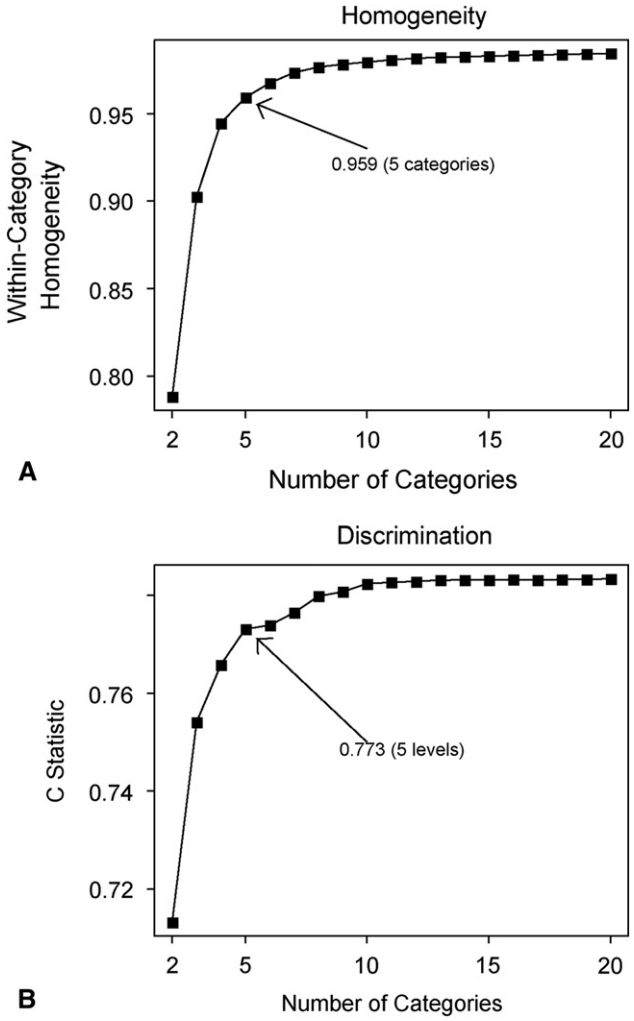

FIGURE 2. Association between number of procedure categories and within-category homogeneity of mortality risk (Panel A) and discrimination for predicting mortality (Panel B). Performance improves with increasing numbers of categories. See Appendix 2 for definition of within-category homogeneity.

Models Combining Scores and Categories With Patient-Level Risk Factors

Two logistic regression models were developed to illustrate the utility of modeling the proposed scores and categories together with patient-level risk factors. The first model included the STS–EACTS score (modeled as a continuous variable) plus 3 patient-level factors: age, weight, and preoperative length of stay. To allow for possible nonlinear effects, the score and the square of the score were both entered in the model. Age and weight were modeled jointly by converting them into a single categorical variable with 7 levels (see Results). Preoperative length of stay was dichotomized as less than or equal to 2 days versus more than 2 days. The second model was identical but used the STS–EACTS categories

TABLE 3. Summary of logistic regression models combining the proposed STS–EACTS scores and categories with patient-level risk factors

Variable	Odds ratio (95% confidence interval)	
	Model 1: STS–EACTS score + patient factors	Model 2: STS–EACTS categories + patient factors
STS–EACTS mortality score		
0.5 vs 0.25	1.4 (1.4–1.5)	–
1.0 vs 0.25	2.6 (2.4–2.8)	–
2.0 vs 0.25	6.3 (5.6–7.1)	–
4.0 vs 0.25	9.4 (8.2–10.8)	–
STS–EACTS mortality category		
Category 1	–	Reference
Category 2	–	2.9 (2.4–3.3)
Category 3	–	4.3 (3.6–5.0)
Category 4	–	7.5 (6.5–8.7)
Category 5	–	15.9 (13.3–18.9)
Age and weight category		
Age ≥1 y	Reference	Reference
Age 1–11 mo, weight ≥6.0 kg	1.0 (0.8–1.2)	0.9 (0.8–1.1)
Age 1–11 mo, weight 4.0–5.9 kg	1.4 (1.2–1.6)	1.3 (1.2–1.5)
Age 1–11 mo, weight <4.0 kg	2.6 (2.2–3.0)	2.6 (2.3–3.0)
Age <1 mo, weight ≥3.0 kg	2.0 (1.8–2.2)	1.9 (1.7–2.2)
Age <1 mo, weight 2.0–2.9 kg	3.3 (2.8–3.8)	3.2 (2.8–3.7)
Age <1 mo, weight <2.0 kg	4.9 (4.2–5.8)	4.9 (4.2–5.7)
Preoperative LOS		
≤2 d	Reference	Reference
>2 d	1.4 (1.3–1.6)	1.4 (1.3–1.5)

STS–EACTS, Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery; *LOS*, length of stay.

(modeled as a set of category indicators) instead of the STS–EACTS score. Additional patient factors, such as comorbidities, were not included because these data were not available to us for the EACTS subset at the time of analysis.

Comparisons With RACHS-1 Categories and ABC Scores

The models described above were also estimated with RACHS-1 categories in place of the STS–EACTS categories and with the ABC score in place of the STS–EACTS score to facilitate comparisons with existing methods. Briefly, the ABC score of a procedure is a number ranging from 1.5 to 15 points that reflects the Aristotle expert panel's assessment of that type of procedure's potential for mortality, morbidity, and technical difficulty. When analyzing operations with multiple procedures, the ABC score was defined as the maximum ABC score across all procedures in the operation. The RACHS-1 methodology divides procedures into 6 categories based on an expert panel's assessment of the procedure's average

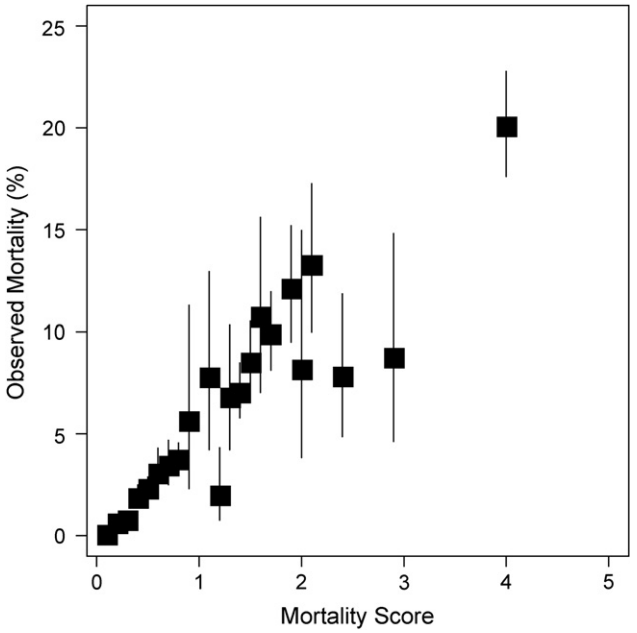

FIGURE 3. Association between Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery score and in-hospital mortality in the validation sample. *Square dots* represent the aggregate mortality rate of procedures sharing the same risk score. Data points with fewer than 40 observations were excluded from the figure. *Vertical lines* represent 95% binomial confidence intervals.

mortality risk, where category 1 has the lowest risk of mortality and category 6 has the highest. Unlike the ABC method, the classification of some procedures is allowed to depend on the patient's age. When analyzing operations with multiple procedures, the operation is assigned to the procedure with the highest RACHS-1 category. Because very few data points

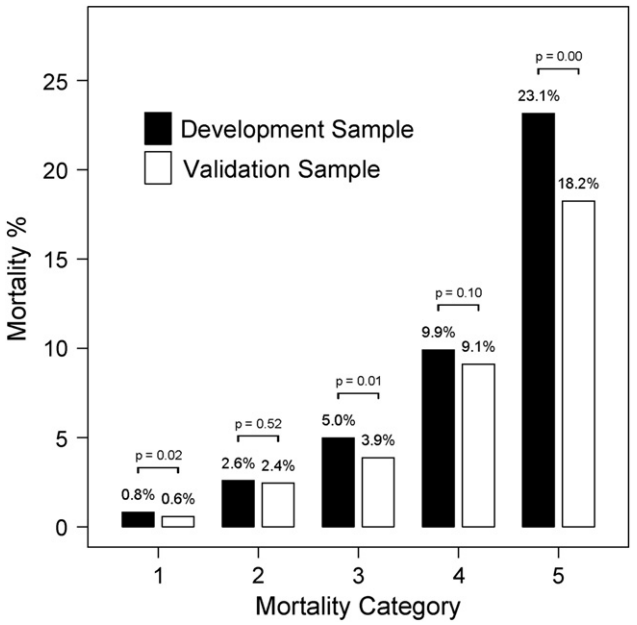

FIGURE 4. Association between proposed risk categories and observed in-hospital mortality.

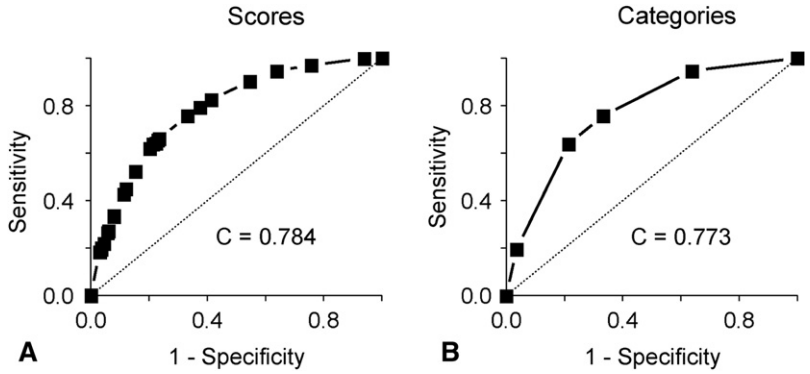

FIGURE 5. Receiver operating characteristic curves for the Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery scores (A) and categories (B) as predictors of in-hospital mortality in the validation sample. The *diagonal line* is provided as a reference. It is the receiver operating characteristic curve that would be observed hypothetically if the scores and categories were not associated with mortality.

were available in RACHS-1 category 5, it was combined with category 6 for analysis. The ''full'' RACHS-1 methodology involves fitting a logistic regression model that includes indicator variables for the RACHS-1 categories together with an indicator variable for single versus multiple cardiac procedures, plus additional adjustment for 3 patient-level risk factors: age, prematurity, and presence of a major noncardiac structural anomaly. Because the required patient-level risk factors were not available in our dataset, we did not implement the full RACHS-1 methodology but instead focused on evaluating the discrimination of the RACHS-1 categories with and without adjustment for patient age, weight, and preoperative length of stay.

Independent Validation Using 2007–2008 Data

The performance of each model was assessed in a separate, more contemporary sample of STS and EACTS data. Overall discrimination was quantified by the C-index. The ability of the proposed score to predict the risk of individual procedures was quantified by calculating the Pearson correlation coefficient between the score and the actual calculated procedure-specific mortality rate in the validation sample. Because sampling variation in the validation sample might artificially increase or decrease the Pearson correlation coefficient, procedures with fewer than 40 occurrences in the validation sample were excluded when calculating the Pearson correlation coefficient. For graphing the association between the proposed score and observed mortality, data from procedures with the same score were aggregated, and the mortality rate of each group of procedures was plotted as a function of the score, excluding groups with fewer than 40 cases. The entire validation was also repeated in the subset of procedures having at least 200 cases in the development sample. Finally, to permit a fair comparison with RACHS-1 and ABC scores, the performance of each model was assessed in the subset of procedures for which both RACHS-1 categories and ABC scores are defined (n = 25,106 patient operations). Statistical comparisons of the C-index for different models were performed using the method of DeLong and colleagues.[11]

TABLE 4. Comparison of C-index for models using the STS–EACTS score, STS–EACTS categories, RACHS-1 categories, and ABC scores*

Method of modeling procedures	Model without patient covariates (C-index)	Model with patient covariates (C-index)
STS–EACTS score	0.787	0.816
STS–EACTS categories	0.778	0.812
RACHS-1 categories	0.745	0.802
ABC score	0.687	0.795

*Validation sample, subset of procedures for which both RACHS-1 categories and ABC scores are defined. *STS–EACTS*, Society of Thoracic Surgeons–European Association for Cardiothoracic Surgery; *RACHS*-1, Risk Adjustment for Congenital Heart Surgery; *ABC*, Aristotle Basic Complexity.

RESULTS

A total of 77,294 patient operations were analyzed, including 3308 (4.3%) in-hospital deaths. There were 71 procedures with at least 200 occurrences, 104 procedures with at least 50 occurrences, and 133 procedures with at least 20 occurrences. Procedures with at least 200 occurrences accounted for 94% of the total patients and 91% of the deaths.

Mortality Rates for Individual Procedures

The frequency of in-hospital mortality for individual procedures ranged from 0% to 40.0%. There were 18 procedures with zero deaths; all of these had sample sizes smaller than 200. When Bayesian modeling was used to estimate mortality risk for individual procedures, the estimates ranged from 0.3% (atrial septal defect repair with patch) to 29.8% (truncus plus interrupted aortic arch repair, Figure 1). For the procedures with more than 200 cases, the raw and model-based estimates were virtually identical (Pearson correlation coefficient > 0.999, Appendix 1).

Mortality Scores and Categories

Names of the procedures analyzed in this study are listed in Table 1, along with their raw and model-based mortality estimates and their proposed scores and categories. The STS–EACTS score takes on values between 0.1 and 5.0 and has 29 unique values. The STS–EACTS categories consist of 5 groups labeled 1 to 5, with higher numbers implying higher mortality risk. The number of patients and procedures per category and their aggregated mortality rates are summarized in Table 2.

The within-category homogeneity criterion and the C-index were plotted as functions of the number of categories to help us determine the optimal number of mortality categories. As shown in Figure 2, *A*, within-category

homogeneity increases rapidly with the number of categories when the number of categories is small. With more than 4 or 5 categories, the homogeneity continues to increase, but the marginal improvement per additional category approaches zero. Similarly, Figure 2, *B*, shows that the estimated discrimination of the categories changes dramatically when the number of groups is varied between 2 and 5, but using more than 5 categories has a relatively modest effect on the C-index. Five categories were chosen as the smallest number that produces both acceptable within-category homogeneity and good discrimination.

Examples of regression models using the proposed scores and categories are summarized in Table 3. The C-index was 0.814 for the model that combined patient factors with the STS–EACTS score and 0.810 for the model that combined patient factors with the STS–EACTS categories. For comparison, when age, weight, and preoperative length of stay were analyzed in a logistic regression model without adjustment for the STS–EACTS scores or categories, the C-index was 0.755.

Validation Using 2007–2008 Data

There was a strong positive association between the proposed STS–EACTS score and actual observed mortality in the validation sample (C-index = 0.784). For the 82 procedures with at least 40 occurrences in the validation sample, the Pearson correlation coefficient between the score of a procedure and its actual observed mortality rate in the validation sample was 0.80. An increasing association between the score and mortality was observed across the range of scores, although several groups of procedures had lower than expected mortality (Figure 3).

The observed mortality rate in the validation sample was slightly lower than in the development sample (3.9% vs 4.3%, $P = .004$), reflecting a trend toward lower mortality in a more contemporary sample. This lower mortality was seen in each of the 5 STS–EACTS categories (Figure 4). Despite the trend toward lower absolute mortality in 2007–2008, the chosen categories continued to perform well at discriminating between high-risk and low-risk procedures (C-index = 0.773). Receiver operating characteristic curves for the proposed scores and categories are displayed in Figure 5. When the validation was repeated in the subset of 73 procedures with at least 200 cases in the development sample, there was a similarly high level of discrimination (C-index = 0.790 for STS–EACTS scores; C-index = 0.782 for STS–EACTS categories) and high correlation between the STS–EACTS score and procedure-specific mortality rates (Pearson correlation coefficient = 0.87).

To assess whether the proposed method discriminates mortality better than the existing RACHS-1 categories and Aristotle scores, each of these was evaluated in the validation sample using the subset of procedures for which both RACHS-1 categories and ABC scores are defined. As summarized in Table 4, discrimination was highest for the STS–EACTS score (C-index = 0.787), followed by the STS–EACTS categories (C-index = 0.778), RACHS-1 categories (C-index = 0.745), and ABC scores (C-index = 0.687, all differences $P < .0001$). Adding patient-level covariates substantially improved each model's discrimination. With the addition of these patient variables, discrimination was highest for the STS–EACTS score (C-index = 0.816), followed by STS–EACTS categories (C-index = 0.812; comparison with STS–EACTS score, $P = .035$), RACHS-1 categories (C-index = 0.802; comparison vs STS–EACTS categories, $P = .008$), and ABC scores (C-index = 0.795; comparison vs STS–EACTS score, $P < .0001$).

DISCUSSION

The goal of this study was to derive a valid tool that can be used to stratify congenital heart surgery procedures based on their relative risk of in-hospital mortality. Using the combined resources of the STS and EACTS databases, we estimated the average mortality rate of 148 procedures and then applied a data-driven algorithm to determine the grouping of procedures that was optimal in the sense of creating internally homogeneous strata. The resulting scores and categories are intended to serve as tools for case-mix adjustment when comparing outcomes of hospitals that perform congenital heart surgery. These measures can be used to perform a stratified analysis that adjusts for type of procedure or they can be included along with patient-level variables in a comprehensive risk adjustment model.

Previous investigators have used a combination of expert opinion and empirical data to group procedures with a similar risk of in-hospital mortality. Experts initially used clinical judgment to group procedures with a similar potential for in-hospital mortality to create the RACHS-1 risk categories. This allocation of procedures was subsequently refined by using empirical data from 2 multi-institutional registries. The goals of the present study were similar to those of RACHS-1 in that we also sought to create internally homogeneous procedure categories using the end point of discharge mortality. A major difference between our approach and the derivation of RAHCS-1 categories is that our procedure categories were determined empirically without the input of an expert panel. When the proposed methodology was assessed in an independent validation sample, models based on the STS–EACTS score and categories had substantially better discrimination than comparable models based on RACHS-1 categories and ABC scores.

Despite the advantages of an empirically based risk stratification system, there are several limitations and caveats.

First, our study focused on estimating procedural mortality and determining homogeneous procedure categories. Additional research is needed to determine the best method of

combining these procedural variables with adjustment for patient-specific risk factors.

Second, despite the large database, several individual procedures had small sample sizes, and the true mortality of these procedures may have been estimated with error. We attempted to minimize this error by using a statistical model, which accounted for small denominators.

Third, because the EACTS and STS registries are voluntary, it is possible that the results observed in this database will differ from those of other nonparticipating institutions.

Fourth, because auditing of the STS and EACTS databases has been limited to a small number of sites, the completeness and accuracy of the data are largely unknown. In an audit of 200 patient records from 10 different STS centers, there was 99.0% agreement in the reporting of discharge mortality by STS sites versus independent auditors and no evidence of selective reporting based on discharge mortality status (personal communication, unpublished STS data).

Another potential limitation rests in the fact that mortality was determined only on the basis of status at the time of discharge. Operative mortality has been defined by the STS Congenital Database Taskforce and the Joint STS–EACTS Congenital Database Committee.[12] It requires knowledge not only of status at discharge but of patient status at 30 days after the operation. Going forward, validation of the STS–EACTS scores and categories using this definition will be possible as the completeness of these data fields in the STS and EACTS databases improves (Appendix 3).

In summary, we have developed a new tool for grouping procedures with a similar empirically estimated risk of in-hospital mortality. Empirically based mortality stratification was possible to a considerable extent because of the large sample sizes of the STS and EACTS congenital databases. The resulting scores and categories can be incorporated into case-mix adjustment methods, such as stratification and regression analysis, to compare institutions on a level playing field.

References

1. 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.
2. 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.
3. Jenkins KJ. Risk adjustment for congenital heart surgery: the RACHS-1 method. *Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu*. 2004;7:180-4.
4. 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.
5. Al-Radi OO, Harrell FE Jr, Caldarone CA, McCrindle BW, Jacobs JP, Williams MG, et al. 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. *J Thorac Cardiovasc Surg*. 2007;133:865-75.
6. Kang N, Tsang VT, Elliott MJ, de Leval MR, Cole TJ. Does the Aristotle score predict outcome in congenital heart surgery? *Eur J Cardiothorac Surg*. 2006; 29:986-8.
7. O'Brien SM, Jacobs JP, Clarke DR, Maruszewski B, Jacobs ML, Walters HL 3rd, et al. Accuracy of the Aristotle Basic Complexity score for classifying the mortality and morbidity potential of congenital heart surgery operations. *Ann Thorac Surg*. 2007;84:2027-37.
8. Jacobs JP, Jacobs ML, Maruszewski B, Lacour-Gayet FG, Clarke DR, Tchervenkov CI, et al. Current status of the European Association for Cardio-Thoracic Surgery and the Society of Thoracic Surgeons Congenital Heart Surgery Database. *Ann Thorac Surg*. 2005;80:2278-84.
9. O'Brien SM. Cutpoint selection for categorizing a continuous predictor. *Biometrics*. 2004;60:504-9.
10. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36.
11. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837-45.
12. Jacobs JP, Mavroudis C, Jacobs ML, Maruszewski B, Tchervenkov CI, Lacour-Gayet FG, et al. What is operative mortality? Defining death in a surgical registry database: a report of the STS Congenital Database Taskforce and the Joint EACTS-STS Congenital Database Committee. *Ann Thorac Surg*. 2006;81: 1937-41.

CHD

Appendix 1. Statistical Model for Estimating Procedure-Specific Mortality Rates

Procedure-specific mortality rates were estimated by using a hierachical (random effects) model. For each of the 148 procedures in the analysis, the number of deaths was modeled by using the following binomial distribution:

$$y_j \sim \text{Binomial}(n_j, \pi_j),\ j = 1, 2, \ldots, 148,$$

where π_j denotes the unknown theoretical probability of mortality for the *j-th* procedure, n_j denotes the number of patients undergoing the procedure in the database (denominator), and y_j denotes the actual observed number of mortalities in the database (numerator). Variation in the theoretical probability of mortality was modeled by assuming the log odds were normally distributed. Thus the model is as follows:

$$\log\left(\pi_j/\left[1-\pi_j\right]\right) = \eta_j;$$

$$\eta_j \overset{\text{ind}}{\sim} N\left(\mu, \sigma^2\right),$$

where μ and σ^2 denote the unknown mean and variance, respectively, of the assumed normal random effects distribution. Parameters of the model were estimated in a Bayesian framework using WinBUGS software. A vague (noninformative) prior distribution was chosen for the parameters μ and σ^2. The

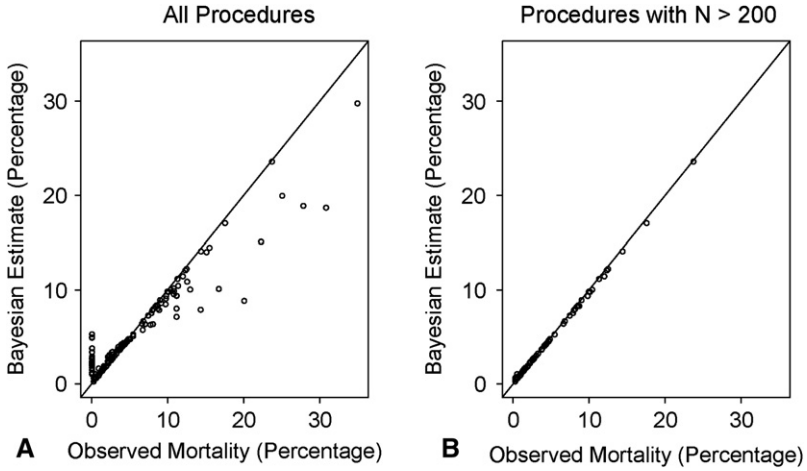

FIGURE 6. Relationship between Bayesian model–based estimates and unadjusted mortality rates for individual procedures in the development sample.

WinBUGS code for this model is available from the authors on request.

As shown in Figure 6, *A*, there was a high degree of correlation between the Bayesian model–based estimate of a procedure's risk and the simple raw unadjusted mortality percentage; however, several procedures had large discrepancies. The difference between the model-based versus raw estimates decreased with increasing sample size. For procedures with more than 200 cases, the raw and model-based estimates were virtually identical (Pearson correlation coefficient > 0.999; Figure 6).

Appendix 2. Methodology for Creating Internally Homogeneous Risk Categories

Procedures were first sorted in order of increasing estimated risk (based on the model in Appendix 1) and then grouped into homogeneous categories to create the risk categories. Let π_i denote the true unknown mortality for the *i-th* procedure, and let $\widehat{\pi}_i$ denote the corresponding estimate. We first sorted procedures so that $\widehat{\pi}_1 < \widehat{\pi}_2 < \cdots < \widehat{\pi}_{148}$. Let k denote the number of categories and let $\boldsymbol{c}_k = \{c_1 < c_2 < \cdots < c_{k-1}\}$ denote a set of category cut points that partition the categories into k groups. The symbol c_j denotes a number between 1 and 148 and represents the index of the highest-risk procedure in the *j-th* category. Also, define $c_0 = 0$ and $c_k = 149$. For any particular choice of k and $\boldsymbol{c}_k$, within-category homogeneity is measured by the weighted sum-of-squares criterion:

$$\mathrm{WSS}(\boldsymbol{c}_k;\pi) = \sum_{j=1}^{k} \sum_{i=c_{j-1}+1}^{c_j} \frac{n_i(\pi_i - \overline{\pi}_j)^2}{\pi_i(1-\pi_i)},$$

where $\overline{\pi}_j = \sum_{i=c_{j-1}+1}^{c_j} n_i\pi_i / \sum_{i=c_{j-1}+1}^{c_j} n_i$ denotes the average risk of mortality among all procedures in the *j-th* category. This criterion is similar to one that has been used previously for defining optimum cut points for categorizing a continuous explanatory variable.[9] The notation $\mathrm{WSS}(\boldsymbol{c}_k;\pi)$ is intended to emphasize that WSS is a function of the chosen cut points $\boldsymbol{c}_k$ and also depends on the unknown procedure-specific probabilities π_i. If the π_i were known instead of unknown, then the ''optimal'' cut points could (in theory) be determined by enumerating all possible choices for the c_j and choosing the one that minimizes the WSS. Because the π_i are unknown, we instead choose cut points that minimize the Bayesian estimate of $\mathrm{WSS}(\boldsymbol{c}_k;\pi)$. Specifically, we chose the cut points that minimize the estimated Bayesian posterior mean as follows:

$$\widehat{\mathrm{WSS}}(\boldsymbol{c}_k) = \frac{1}{3000}\sum_{h=1}^{3000} \mathrm{WSS}\left(\boldsymbol{c}_k;\pi^{(h)}\right),$$

where $\pi^{(h)}$ denotes a random draw from the joint posterior distribution of the π_i's. Finding the set of cut points that minimizes this quantity exactly is technically challenging and required the use of a novel dynamic programming algorithm (unpublished).

The criterion described above gets smaller as the within-category homogeneity improves. For plotting the change in homogeneity versus k, it is intuitively appealing to use a criterion that increases rather than decreases. The criterion used in Figure 2 (and throughout the article) is defined as follows:

$$\text{Homogeneity} = 1 - \widehat{\mathrm{WSS}}(\boldsymbol{c}_k)/\widehat{\mathrm{WSS}}(\boldsymbol{c}_1).$$

This criterion ranges from 0.0 to 1.0 and increases as the categories become more homogeneous.

Appendix 3. Completeness of STS Mortality Data

The mortality end point for this study was mortality status at the time of discharge, ie, in-hospital mortality. It was chosen over operative mortality (ie, death prior to discharge or after discharge but within 30 days of surgery) or 30-day mortality status in large part because 30-day status is frequently missing whereas discharge mortality is rarely missing. As shown in Figure 7, the completeness of 30-day mortality status has improved over time. In the future, it may be feasible to adapt the STS-EACTS methodology (or develop a new methodology) to predict the endpoint of operative mortality or 30-day mortality, assuming the completeness of 30-day mortality reporting continues to improve.

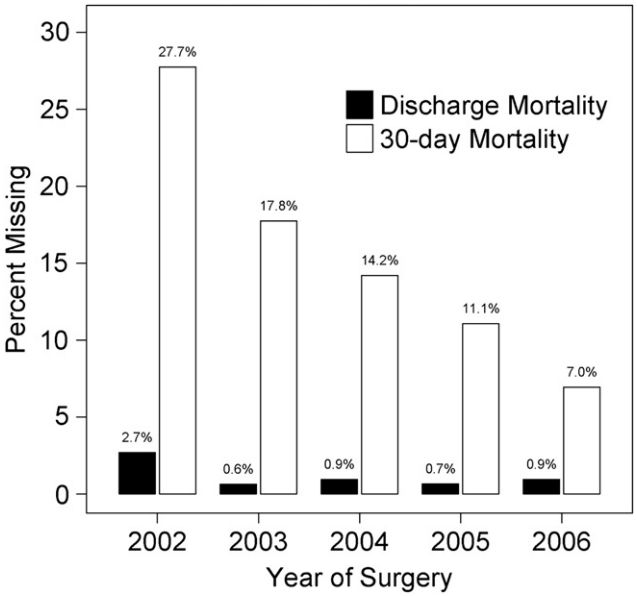

FIGURE 7. Decreasing percentage of missing data in the fields ''mortality discharge status'' (alive or dead) and ''status at 30 days after surgery'' (alive, dead, or unknown) in the Society of Thoracic Surgeons Congenital Database from 2002 to 2006.

CHD