Published in final edited form as:
J Am Coll Surg. 2009 November ; 209(5): 551–556. doi:10.1016/j.jamcollsurg.2009.08.008.

Predicted Risk of Mortality Models: Surgeons Need to Understand Limitations of the University Health System Consortium Models

Benjamin D Kozower, MD, MPH, FACS, **Gorav Ailawadi, MD, FACS**, **David R Jones, MD, FACS**, **Robert D Pates, PhD**, **Christine L Lau, MD, FACS**, **Irving L Kron, MD, FACS**, and **George J Stukenborg, PhD**
Departments of Surgery (Kozower, Ailawadi, Jones, Lau, Kron) and Public Health Sciences (Pates, Stukenborg), University of Virginia Health System, Charlottesville, VA.

Abstract

BACKGROUND—The University HealthSystem Consortium (UHC) mortality risk adjustment models are increasingly being used as benchmarks for quality assessment. But these administrative database models may include postoperative complications in their adjustments for preoperative risk. The purpose of this study was to compare the performance of the UHC with the Society of Thoracic Surgeons (STS) risk-adjusted mortality models for adult cardiac surgery and evaluate the contribution of postoperative complications on model performance.

STUDY DESIGN—We identified adult cardiac surgery patients with mortality risk estimates in both the UHC and Society of Thoracic Surgeons databases. We compared the predictive performance and calibration of estimates from both models. We then reestimated both models using only patients without any postoperative complications to determine the relative contribution of adjustments for postoperative events on model performance.

RESULTS—In the study population of 2,171 patients, the UHC model explained more variability (27% versus 13%, $p < 0.001$) and achieved better discrimination (C statistic = 0.88 versus 0.81, $p < 0.001$). But when applied in the population of patients without complications, the UHC model performance declined severely. The C statistic decreased from 0.88 to 0.49, a level of discrimination equivalent to random chance. The discrimination of the Society of Thoracic Surgeons model was unchanged (C statistic of 0.79 versus 0.81).

CONCLUSIONS—Although the UHC model demonstrated better performance in the total study population, this difference in performance reflects adjustments for conditions that are postoperative complications. The current UHC models should not be used for quality benchmarks.

Surgeons and hospitals face an increasing demand to provide evidence for the quality of care they deliver. Reporting the outcomes of surgical procedures as a proxy measurement of quality has been routine practice for cardiac surgery for over a decade.[1,2] Mortality rates

Correspondence address: Benjamin D Kozower, MD, MPH, University of Virginia Health System, General Thoracic Surgery, PO Box 800679, Charlottesville, VA 22908-0679.

Disclosure Information: Nothing to disclose.

Author Contributions
Study conception and design: Kozower, Ailawadi, Stukenborg
Acquisition of data: Kozower, Pates
Analysis and interpretation of data: Kozower, Jones, Lau, Kron, Stukenborg
Drafting of manuscript: Kozower, Ailawadi, Lau, Stukenborg
Critical revision: Jones, Pates, Kron

vary across providers and hospitals and the presumption is that, after controlling for patients' severity of illness, residual differences in mortality rates relate to differences in the quality of care.[3,4] Risk adjustment methodologies have been developed to specifically address these concerns. More than a dozen risk adjustment tools are available using either clinical measures from clinical databases or code-based measures from administrative databases.

The University HealthSystem Consortium (UHC) was formed in 1984. It's an alliance of 101 academic medical centers and nearly 200 of their affiliate hospitals, representing more than 90% of the nation's nonprofit academic medical centers.[5] The UHC risk-adjusted mortality models are increasingly used as a benchmark for quality assessment across its member institutions, and our institution recently considered using their models as the primary benchmark for surgical quality. The UHC models are based on discharge abstracts and include adjustments for differences in patient severity using the All Patient Refined Diagnosis Related Groups (APR-DRG), developed by 3M Health Information Systems.[6]

The thoracic surgery community has long recognized the importance of risk adjustment and predictive modeling. The Society of Thoracic Surgeons (STS) cardiac database is a clinical database created in 1986. The STS database has revolutionized the ability of cardiac surgeons to risk adjust outcomes and modify clinical practice patterns.[7,8] The purpose of this study was to compare performance of the UHC with the STS risk-adjusted mortality models for adult cardiac surgery and evaluate the contribution of postoperative complications on model performance.

METHODS

Patient selection

We identified all adult patients undergoing cardiac surgery at the University of Virginia between January 2003 and January 2008, with records in the STS database, who underwent any of the following procedures: coronary artery bypass grafting (CABG) only, aortic valve replacement only, mitral valve replacement only, aortic valve replacement + coronary artery bypass grafting, and mitral valve replacement + coronary artery bypass grafting. These patients with STS mortality risk scores were then matched to the UHC database to identify patients with both STS and UHC mortality risk scores. Observed perioperative mortality was identified for each patient, defined as death within 30 days of operation or within the same hospitalization. Postoperative complications, overall and by specific type, were identified for each patient as reported in the STS database.

STS data is acquired by the University of Virginia Heart and Vascular Center and submitted to the STS quarterly. A clinical/research nurse prepares the data abstracts for the STS and runs extensive audits on the data for consistency and completeness. Internal audits of the data demonstrated that it is more than 99% complete, because we looked at every metric for every cardiac case performed. The Health Services/Computer Services Decision Support team sends data to the UHC on a monthly basis. The source of this data is the financial and administrative data from the University of Virginia Health System. The ICD-9-CM codes and the demographic information reported to the UHC are drawn from the same data source used to report patient billing data to Medicare and other payers.

Statistical analysis

Logistic regression analysis was used to calculate the probability of perioperative death for each patient in the study population using only the STS mortality risk score as a predictor of observed mortality. Logistic regression analysis was similarly used to calculate the probability of perioperative death using only the UHC mortality risk score. The statistical

performance obtained by each model was assessed by calculating the maximum adjusted R^2 and the C statistic.[9,10] The maximum adjusted R^2 statistic measures the proportion of the log likelihood explained by the model compared with that obtainable by a perfect model. The C statistic, which is equivalent to the area under the receiver operating characteristics curve (AUC), measures the discrimination accuracy of the models.[11] A C statistic value of 0.5 indicates that the model is equivalent to random chance; a value of 1.0 indicates that the model achieves perfect discrimination between survivors and decedents.

Patients in the study population were ranked into deciles by their predicted probability of death, using the mortality risk estimated by the STS risk adjustment model alone and also using the UHC risk adjustment model alone. Model calibration was assessed by comparing the observed number of deaths to the sum of the probabilities of death for patients in each decile of predicted risk. The statistical significance of the difference in the observed-to-expected numbers of deaths for patients across deciles within each model was assessed using the Hosmer-Lemeshow chi-squared test statistic (with eight degrees of freedom).

The extent of the difference in how individual patients were ranked using the STS and UHC scores was also examined. Differences in the decile of predicted risk assigned to individual patients between models were assessed by cross-tabulation of the results from the two models. The proportion of patients who differed by two or more deciles of risk between models was calculated as a measure of model agreement, as performed by Iezzoni and others.[9]

The statistical significance of the difference in predictive information contributed by the STS and UHC scores was assessed using multivariable logistic regression analysis. A model was estimated that included both the STS and UHC scores as predictors of perioperative death. In this combined model, the Wald chi-squared test statistic provides a specific test of the statistical significance of the independent effect of the additional covariate, adjusted for the contribution of the other covariate.

The mortality risk scores included as covariates in this analysis are probabilities of death obtained directly from the STS and UHC. These risk scores were calculated with fixed multivariable equations applied to each patient in the study population, using information from detailed clinical and administrative data that were supplied to the STS and UHC. The actual model parameters used to estimate the STS and UHC risk scores are unknown in this analysis.

Contribution of postoperative complications

We were interested in assessing if information from postoperative complications contributed to the performance of the models. To address this question, the original study population was divided into two groups: patients without any complications, and patients with any complication. Complications were identified in the STS database and defined using the definitions from database version 2.61.[12] Both logistic regression models were then reestimated in each of the two subpopulations. Comparison of model statistical performance obtained in the original population with that in the population without any complications provides an empiric test of the contribution made by information from postoperative complications.

All analyses were performed using SAS 9.1.3 software. The Human Studies Committee at the University of Virginia granted approval for this research and waived the need for individual consent.

RESULTS

A total of 2,171 cardiac surgery patients were identified with mortality risk scores available in both the STS and UHC databases. The overall mortality rate was 2.7% (58 of 2,171). The mean age of the study population was 65.7 years, and men accounted for 73.6% (1,598 of 2,171). Table 1 lists the frequency of each cardiac surgery procedure performed in the study population by type and age group distribution. Complications were identified in 29.6% of the study population (643 of 2,171) and are listed in Table 2 according to the STS database definitions for postoperative complications, version 2.61.[13]

Table 3 presents a summary of the logistic model results for the total study population. These results demonstrate that the STS risk-adjusted mortality model explained 13% of the log-likelihood of mortality (maximum adjusted R^2) obtainable; the UHC model explained 27% ($p < 0.001$). The UHC model also achieved better discrimination between survivors and decedents in the total study population than the STS model (C statistic = 0.88 versus 0.81). The Wald chi-squared test statistics obtained by including both the STS and UHC scores as covariates in the same model demonstrated that the difference in predictive information between the STS and UHC scores was statistically significant ($p < 0.0001$).

Both models demonstrated statistically significant differences in the calibration of predictions across ranges of risk. Hosmer-Lemeshow test statistic p values for the STS and UHC risk score models were 0.01 and 0.03, respectively. The model using the UHC score demonstrated a much larger discrepancy for patients at the highest level of risk. Substantial differences occurred between the models in the estimated mortality risk for individual patients. Although 72.7% (1,579 of 2,171) of patients were ranked within two deciles by the STS and UHC models, the STS model ranked 13.9% (303 of 2,171) of patients at least two deciles of risk higher and 13.31% (289 of 2,171) of patients at least two deciles of risk lower. Scores from the two models tended to agree most for patients at the extremes of mortality risk. Figure 1 provides a bubble plot of the specific distribution of the agreement and disagreement between models across deciles. Each bubble is sized to depict the relative total proportion of the study population matched by deciles of predicted mortality between the STS and UHC scores.

Adjustment for postoperative complications

The statistical performance of the logistic model using the UHC score declined severely when applied in the population of patients without complications. The proportion of variability explained by the model using the UHC score was reduced from 27.3% to 1.0%, and the level of discrimination obtained between survivors and decedents declined from 0.88 to 0.49, a level nearly equivalent to random chance. In contrast, the model using the STS score achieved nearly the same level of discrimination obtained in the total study population (0.79 versus 0.81 in the total population). Although the proportion of variability explained by the STS declined from 12.9% to 5.3%, the relative decline was much less than that demonstrated for the model using the UHC score. Both models, applied in this patient population without complications, obtained good calibration across deciles of mortality risk (Hosmer-Lemeshow test statistic p values of 0.45 and 0.24), which, in part reflects the small number of deaths in this subpopulation (0.5%).

DISCUSSION

Risk-adjusted mortality is an important component of physician and hospital “report cards” and a common surrogate for quality. Because of the importance of these measures, surgeons should play an active role in evaluating the validity of these models. The UHC risk-adjusted mortality model is based on abstracted data from hospital discharge records. This type of

data is appealing because it's readily available, computerized, and relatively inexpensive.[14] The UHC models are also appealing to academic medical centers because they appear to provide a convenient way to benchmark their performance against similar institutions.

The UHC administrative database model appears significantly better at predicting death than the STS clinical database model (measured by R^2 value and C statistic) in the total study population of patients undergoing cardiac surgery. These results are consistent with those from other studies, demonstrating that predictive models using administrative data have better predictive performance than models using clinical data.[9,14] But the improved predictive ability of the UHC model is a reflection of its inclusion of postoperative complications in its preoperative risk model. By reestimating the performance of the UHC model in a subpopulation that specifically excludes patients with complications, we demonstrated that the UHC model's statistical performance falls off sharply. Importantly, the STS model applied to the same patients without complications demonstrates much less attenuation in performance.

The UHC model includes adjustments using the APR-DRG system created by the 3M Health Information Systems.[6] The system's vendor claims that APR-DRGs are "the most comprehensive, clinically accurate severity of illness and risk of mortality product available." In an effort to distinguish between preoperative and postoperative events, the APR-DRG grouper assigns the lowest severity of illness level and risk of mortality level to ICD-9 complication codes to limit their contribution to mortality risk. But certain codes such as stroke and renal failure are difficult to distinguish as preoperative or postoperative without a date associated with the diagnosis code.

The UHC risk-adjusted mortality model, like other APR-DRG-based administrative models, can potentially include conditions that are postoperative complications. This can happen because the UHC model includes all discharge codes for conditions included in the adjustment, regardless of when these events occurred. So, complications including postoperative stroke, renal failure or cardiac arrest may be included in the adjustment and actually increase a patient's preoperative risk.

When comparing model agreement, the STS model ranked patients at least 2 deciles of risk higher than the UHC model in 13.9% of patients and at least 2 deciles of risk lower in 13.3% of patients. So the difference in statistical performance matters, because the relative order of patients ranked by mortality risk is substantially different for a large proportion of the total patient population, depending on which risk adjustment method is used.

There are several limitations to this study. First, it is a retrospective comparison of data from a single academic referral center, and this patient population may differ from those in other cardiac surgical units. Second, the variables and their weights used in the UHC predicated mortality algorithm are proprietary and were not available for this research. Sowe could not compare them directly to the STS model. Third, this analysis does not directly prove that the UHC model performs better because it incorrectly classifies postoperative complications as preoperative risk factors. But this conclusion is supported by our empirical test of the models both with and without patients having postoperative complications. In addition, a similar finding was demonstrated by Romano and Chan[15] when risk adjusting acute myocardial infarction mortality. When trained blinded coders reabstracted data to establish the timing of each diagnosis code in patients with acute myocardial infarction, the APR-DRG model performance decreased significantly when postadmission diagnoses were excluded from the risk adjustment.

Steinberg and colleagues[16] recently reported a comparison of the UHC and National Surgical Quality Improvement Program risk-adjustment methodologies in surgical quality

improvement. They discovered that the risk adjustment method used had a dramatic impact on the quality assessment of their division and that significant differences in reporting of both comorbidities and outcomes were likely responsible. Unlike our study, Steinberg and colleagues[16] did not assess the temporal relationship of comorbidities and outcomes to determine its contribution to their findings.

Recent movement in pay-for-performance efforts has influenced the need to include present-on-admission (POA) indicators on hospital discharge codes (DRGs).[17] Accordingly, the DRG payment determination should include only diagnoses present on admission and exclude conditions that originate during the hospital stay.[18] Beginning October 1, 2008, the Centers for Medicare and Medicaid Services required hospitals to report a POA indicator for each ICD-9 diagnosis code listed for a hospitalization.[13] The POA indicator will enable the UHC and other administrative models to better distinguish between preexisting conditions and complications.[5]

The current UHC risk-adjusted mortality models are not appropriate for preoperative quality assessment. They are appropriate, as demonstrated by our findings, only if the purpose is to adjust for both comorbidities and complications. Although POA coding will make a substantial improvement in the measurement of baseline risk, it will not solve all of the problems with the UHC or other models relying on administrative data. For example, diagnosis codes for situations such as an intraaortic balloon pump inserted at a different hospital before transferring a patient, will not be accounted for in the APR-DRG system. The importance of this will vary significantly depending on the referral and transfer patterns of a hospital. In their current form, the UHC risk-adjusted mortality models inappropriately adjust for postoperative complications and are not suitable for use as a quality metric.

Abbreviations and Acronyms

APR-DRG	All Patient Refined Diagnosis Related Group
POA	present on admission
STS	Society of Thoracic Surgeons
UHC	University HealthSystem Consortium

REFERENCES

1. Hannan EL, Racz MJ, Jollis JG, Peterson ED. Using Medicare claims data to assess provider quality for CABG surgery: Does it work well enough? Health Serv Res. 1997; 31:659–678. [PubMed: 9018210]
2. Griffith BP, Hattler BG, Hardesty RL, et al. The need for accurate risk-adjusted measures of outcome in surgery. Lessons learned through coronary artery bypass. Ann Surg. 1995; 222:593–598. [PubMed: 7574937]
3. Hannan EL, Kilburn H Jr, O'Donnell JF, et al. Adult open heart surgery in New York State. An analysis of risk factors and hospital mortality rates. JAMA. 1990; 264:2768–2774. [PubMed: 2232064]
4. Chassin MR, Hannan EL, DeBuono BA. Benefits and hazards of reporting medical outcomes publicly. N Engl J Med. 1996; 334:394–398. [PubMed: 8538714]
5. University HealthSystem Consortium. [Accessed December 30, 2008] Available at: http://www.uhc.edu/12443.htm.
6. 3M Health Information Systems. , editor. All Patient Refined Diagnosis Related Groups. Definition Manual. Wallingford, CT: 3M Health Information Systems; 2005. p. 993

7. Clark RE. The development of the Society of Thoracic Surgeons voluntary national database system: Genesis, issues, growth, and status. Best Practices & Benchmarking in Healthcare: a Practical Journal for Clinical & Management Applications. 1996; 1:62–69.
8. Mack MJ, Herbert M, Prince S, et al. Does reporting of coronary artery bypass grafting from administrative databases accurately reflect actual clinical outcomes? J Thorac Cardiovasc Surg. 2005; 129:1309–1317. [PubMed: 15942571]
9. Iezzoni LI, Ash AS, Shwartz M, et al. Predicting who dies depends on how severity is measured: Implications for evaluating patient outcomes. Ann Intern Med. 1995; 123:763–770. [PubMed: 7574194]
10. Iezzoni LI, Shwartz M, Ash AS, et al. Risk adjustment methods can affect perceptions of outcomes. Am J Med Qual. 1994; 9:43–48. [PubMed: 8044051]
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]
12. [Accessed March 17, 2009] The Society of Thoracic Surgeons – Data Collection Version 2.61. Available at: http://www.sts.org/sections/stsnationaldatabase/datamanagers/adultcardiacdb/datacollection/index.html.
13. Centers for Medicare and Medicaid Services. [Accessed January 06, 2009] Present on admission indicator. Available at: http://www.cms.hhs.gov/HospitalAcqCond/01_Overview.asp#TopOfPage.
14. Iezzoni LI, Ash AS, Shwartz M, et al. Predicting in-hospital deaths from coronary artery bypass graft surgery. Do different severity measures give different predictions? [see comment]. Med Care. 1998; 36:28–39. [PubMed: 9431329]
15. Romano PS, Chan BK. Risk-adjusting acute myocardial infarction mortality: Are APR-DRGs the right tool? [see comment]. Health Serv Res. 2000; 34:1469–1489. [PubMed: 10737448]
16. Steinberg SM, Popa MR, Michalek JA, et al. Comparison of risk adjustment methodologies in surgical quality improvement. Surgery. 2008; 144:662–667. [PubMed: 18847652]
17. Zhan C, Elixhauser A, Friedman B, et al. Modifying DRG-PPS to include only diagnoses present on admission: Financial implications and challenges. Med Care. 2007; 45:288–291. [PubMed: 17496711]
18. Medicare Payment Advisory Commission. Report to the Congress: Medicare payment policy. Washington, DC: 2005.

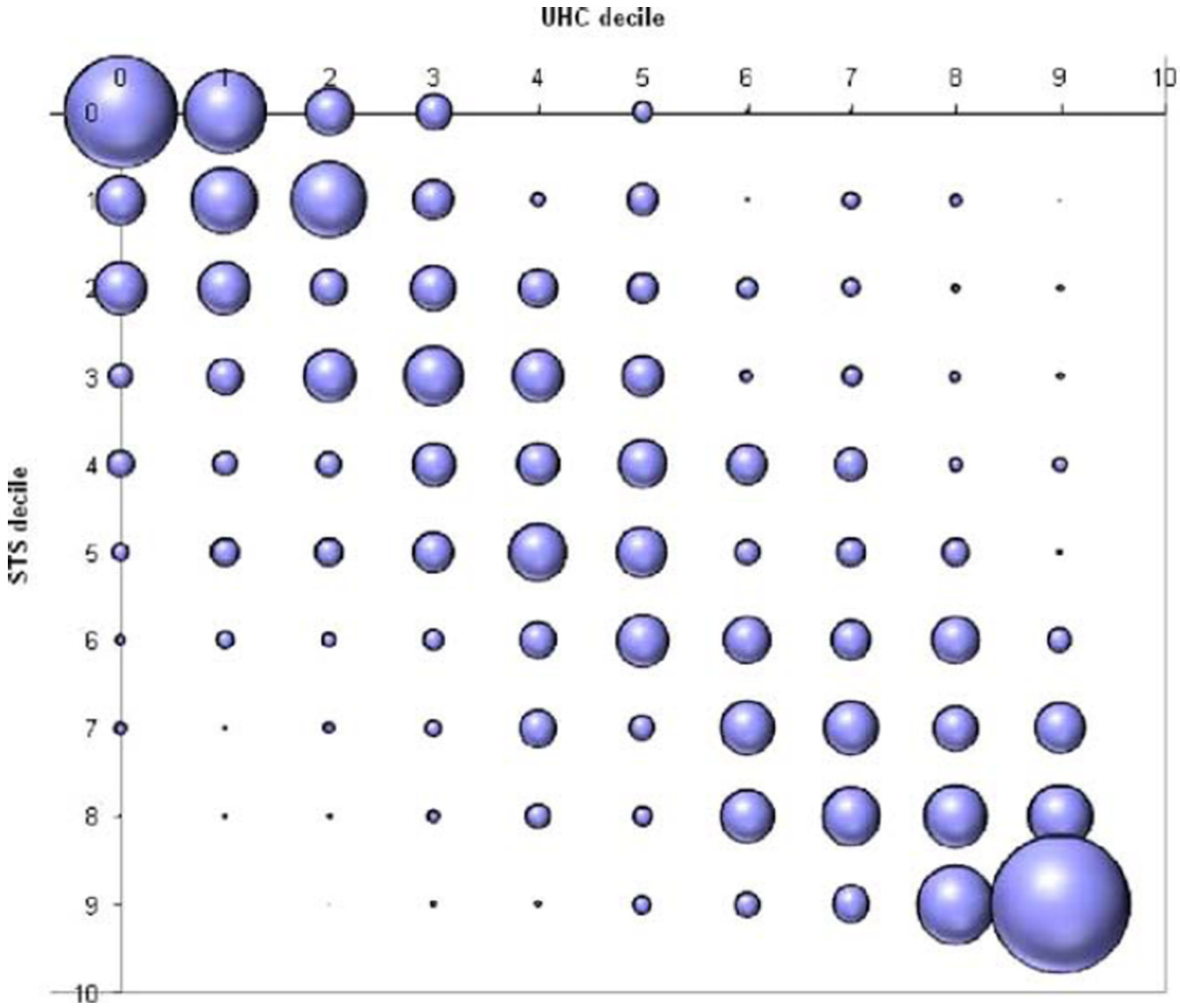

Figure 1.

Proportion of patients matched by deciles of predicted mortality. Between STS and UHC scores; the bubble plot of the specific distribution of the agreement and disagreement between models across deciles. The UHC and STS deciles of predicted mortality risk are shown on the X and Y axis, respectively. Each bubble is sized to depict the relative total proportion of the study population matched by deciles of predicted mortality between the STS and UHC scores. STS, Society of Thoracic Surgeons; UHC, University HealthSystem Consortium.

Table 1

Study Population Characteristics (n = 2,171)

Characteristic	n	Total study population, %
Procedure type		
CABG only	1,601	73.74
AV replacement	282	12.99
AV replacement + CABG	199	9.17
MV replacement	60	2.76
MV replacement + CABG	29	1.34
Age group, y		
20–29	5	0.23
30–39	31	1.43
40–49	164	7.55
50–59	468	21.56
60–69	676	31.14
70–79	609	28.05
80–89	210	9.67
90+	8	0.37

AV, aortic valve; CABG, coronary artery bypass grafting; MV, mitral valve.

Table 2

Postoperative Complications

Complication	n	Total study population, %	Patients with complications, %
Any complication	643	29.62	100.00
Infection, leg	2	0.10	0.31
Infection, septicemia	53	2.45	8.24
Infection, deep sternal	11	0.51	1.71
Neurologic, coma	3	0.14	0.47
Stroke, permanent	38	1.76	5.91
Stroke, transient	4	0.19	0.62
Perioperative myocardial infarction	5	0.24	0.78
Reoperation for tamponade	36	1.66	5.60
Reoperation, other cardiac cause	24	1.11	3.73
Reoperation, other noncardiac cause	60	2.77	9.33
Atrial fibrillation	398	18.34	61.90
Anticoagulation event	11	0.51	1.71
Cardiac arrest	48	2.22	7.47
Gastrointestinal event	35	1.62	5.44
Heart block	25	1.16	3.89
Multisystem organ failure	36	1.66	5.60
Other	2	0.10	0.31
Cardiac tamponade	1	0.05	0.16
Pneumonia	85	3.92	13.22
Pulmonary embolism	3	0.14	0.47
Prolonged ventilation	206	9.49	32.04
Renal failure, dialysis required	50	2.31	7.78
Renal failure, without dialysis	72	3.32	11.20
Acute limb ischemia	9	0.42	1.40
Arterial dissection, iliac or femoral	1	0.05	0.16

Postoperative complications are defined using The Society of Thoracic Surgeons Database version 2.61.[17]

Table 3

Comparison of Model Statistical Performance and Calibration

Variable	Logistic model using STS score	Logistic model using UHC score
Results for total study population, n	2,171	2,171
Observed number of deaths, n (%)	58 (2.7)	58 (2.7)
Wald chi-squared p value for covariate	< 0.0001	< 0.0001
Statistical performance measures		
Maximum adjusted R^2	0.129	0.273
C statistic	0.812	0.885
Calibration by deciles of predicted risk		
Observed/expected deaths in decile 1	0/3.1	0/2.8
Observed/expected deaths in decile 2	1/3.2	0/2.8
Observed/expected deaths in decile 3	1/3.3	1/2.8
Observed/expected deaths in decile 4	1/3.4	0/2.9
Observed/expected deaths in decile 5	2/3.6	1/2.9
Observed/expected deaths in decile 6	2/3.8	1/2.9
Observed/expected deaths in decile 7	5/4.2	4/3.0
Observed/expected deaths in decile 8	11/4.9	3/3.2
Observed/expected deaths in decile 9	10/6.3	6/3.8
Observed/expected deaths in decile 10	25/22.2	42/30.9
Hosmer-Lemeshow test p value (df = 8)	0.010	0.033
Results for patients without any complication, n	1,528	1,528
Observed number of deaths, n (%)	7 (0.5)	7 (0.5)
Wald chi-squared p value	0.006	0.745
Maximum adjusted R^2	0.053	0.001
C statistic	0.794	0.494
Hosmer-Lemeshow test p value (df = 8)	0.446	0.240
Results for patients with any complication, n	643	643
Observed number of deaths, n (%)	51 (7.9)	51 (7.9)
Wald chi-square p value	<0.0001	< 0.0001
Maximum adjusted R^2	0.108	0.283
C statistic	0.739	0.876
Hosmer-Lemeshow test p value (df = 8)	0.086	0.005

STS, Society of Thoracic Surgeons; UHC, University HealthSystem Consortium.