

S.4. – S.11. Measure Specifications

Due to the complex methodology used to construct the composite measure, it is impractical to separately discuss the numerator and denominator. The following discussion describes how each domain score is calculated and how these are combined into an overall composite score.

The STS CABG Composite Score comprises four domains consisting of eleven individual measures:

1. Absence of Operative Mortality
0119 Risk-Adjusted Operative Mortality for CABG
2. Absence of Major Morbidity, scored any-or-none
0131 Risk-Adjusted Postoperative Stroke/Cerebrovascular Accident
0115 Risk-Adjusted Postoperative Surgical Re-exploration
0130 Risk-Adjusted Postoperative Deep Sternal Wound Infection
0114 Risk-Adjusted Postoperative Renal Failure
0129 Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
3. Use of Internal Mammary Artery (IMA)
0134 Use of IMA in CABG
4. Use of All Evidence-based Perioperative Medications, scored all-or-none
0127 Preoperative Beta Blockade
0117 Beta Blockade at Discharge
0116 Anti-Platelet Medication at Discharge
0118 Anti-Lipid Treatment Discharge

Participants receive a score for each of the four domains, plus an overall composite score. The overall composite score is created by “rolling up” the four domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by one star (below average performance), two stars (average performance), or three stars (above average performance).

Patient Population: The analysis population consists of patients aged 18 years or older who undergo isolated CABG surgery

Time Period: 12 months

Data Completeness Requirement: Participants are excluded from the analysis if they have fewer than 10 isolated CABG procedures in the patient population or if they have more than 5% missing data on any of the previously mentioned five NQF-endorsed process measures.

Technical Details

The unit of measurement for the STS CABG Composite Score can be either a participant (most often a cardiac surgical practice but occasionally an individual surgeon) or a hospital.

Each domain score has a theoretical range of 0 to 1 and is interpreted as a probability. A description of these probabilities is presented in the table below. Larger values imply better performance. Although the theoretical range of each score (probability) is 0 to 1, the actual scores tend to be clustered in the upper end of the 0-1 interval. For reporting purposes, the probabilities are expressed as percentages ranging from 0% to 100%.

#	Domain	Interpretation of Domain Score
1	Absence of Operative Mortality	π_1 = The probability (risk-adjusted) that a patient will be discharged alive and will survive to >30 days post-surgery.
2	Absence of Major Morbidity	π_2 = The probability (risk-adjusted) that a patient will be discharged without experiencing <u>any</u> of the following endpoints: stroke/cerebrovascular accident, surgical re-exploration, deep sternal wound infection, post-operative renal failure, prolonged intubation

		(ventilation).
3	Use of Internal Mammary Artery (IMA)	π_3 =The probability that a patient without a prior CABG will receive an IMA. Note: Patients with prior CABG surgery or with documented contraindication for IMA use (subclavian stenosis, previous cardiac or thoracic surgery, previous mediastinal radiation, an emergent or salvage procedure or no LAD disease) are not included in the denominator
4	Use of All Evidence-based Perioperative Medications	π_4 =The probability that a patient will receive <u>all</u> of the medications for which the patient is eligible from the following list: preoperative beta blockade; discharge beta blockade, antiplatelet agents, antilipid agents. Note: Discharge medications are not required for patients who died prior to discharge.

Separate probability estimates are calculated for each unit in the analysis. The method of Bayesian multivariate hierarchical regression modeling is used to obtain estimates that account for chance variation and noisy data. The Bayesian statistical framework is used to assess statistical significance, calculate measures of uncertainty, and distinguish true variation from random noise. After estimating the probability parameters for a unit, a composite score is calculated for each unit by using the following formula:

$$\text{STS composite score} = w_1 \times \hat{\pi}_1 + w_2 \times \hat{\pi}_2 + w_3 \times \hat{\pi}_3 + w_4 \times \hat{\pi}_4$$

where $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\pi}_4$ denote the estimate (for the unit of interest) of the probabilities defined in the table above and w_1, w_2, w_3, w_4 denote the weights applied to the four domains. The weights were derived during original measure development to reflect the relative size of variation as well as relative clinical importance of the four domains.

Star Rating: Star ratings are derived by testing whether the participant's composite or domain score is significantly different from the overall STS average. For instance, if for each of the 4 composite score domains, a participant's estimated score is lower than the overall STS average, but the difference between the participant and STS is not statistically significant, the ratings would each be 2 stars. If however, for the overall composite, the point estimate is lower than the STS average, AND this difference is statistically significant, the overall participant star rating is 1 star. The fact that statistical significance is achieved for the composite score but not the individual domains reflects the greater precision of the composite score compared to individual endpoints. This precision is achieved by aggregating information across multiple endpoints instead of a single endpoint.

The current version of the STS CABG risk models can be found in the following article:

Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 1—Coronary Artery Bypass Grafting Surgery. *Ann Thorac Surg* 2009;88:S2-S22.

Additional details regarding the CABG Composite Score are provided in the attached manuscript:

O'Brien SM, Shahian DM, DeLong ER, Normand SL et al. Quality measurement in adult cardiac surgery: part 2--Statistical considerations in composite measure scoring and provider rating. *Ann Thorac Surg* 2007;83(4):S13-26.

SQL code to create function to identify procedures.txt

BEGIN

```
-- Start by identifying the cases where procedures were performed that definitively put the case into the
Other category. ProcID=null.
  if (VSTCV=1 or EndoProc=1 or OCarACDLE=1 or ResectSubA=1 or OCarCrTx=1 or OCarSVR=1 or CCancCase=1) or
(OCTumor<>1 and OCTumor is not null) or (OCPulThromDis<>1 and OCPulThromDis is not null) then
    Return null;
  else
    if (VADProc=2 and (UnplVAD=2 or UnplVAD is null)) or VADProc=3 or VADProc=4 then
      Return null;
    else
      if OCarASD=1 and (OCarASDTy=1 or OCarASDTy=2 or OCarASDTy is null) then
        Return null;
      else
        if OCarAFibSur=1 and OCarAFibAProc=2 then
          Return null;
        else
          if (OpTricus is not null and OpTricus<>1) or (OpPulm is not null and OpPulm<>1) then
            if UnplProc=1 or UnplProc=2 or UnplProc is null then
              Return null;
            else
              if UnplCABG=1 or UnplAV=1 or UnplMV=1 or UnplAo=1 or UnplVAD=1 then
                Return null;
              end if;
            end if;
          end if;
          if (UnplOth=2 or UnplOth is null) or UnplProc=2 then
            if OpONCard=1 or OCarLVA=1 or OCarVSD=1 or OCarTrma=1 or OCarOthr=1 then
              Return null;
            end if;
          end if;
          if (OCAoProcType is not null and OCAoProcType<>1) then
            if (UnplAo=2 or UnplAo is null) or (UnplAo=1 and UnplProc=2) then
              Return null;
            end if;
          end if;
        end if;
      end if;
    end if;
  end if;
end if;
```

SQL code to create function to identify procedures.txt

```
-- Now determine whether the procedure is an isolated CAB. ProcID=1.
if OpCAB=1 and (UnplCABG=2 or UnplCABG is null) then
    if OpValve=2 or OpValve is null then
        if (OCarCongProc1 is null or OCarCongProc1=10 or OCarCongProc1=1291 or OCarCongProc1=1305) and
            (OCarCongProc2 is null or OCarCongProc2=10 or OCarCongProc2=1291 or
OCarCongProc2=1305) and
            (OCarCongProc3 is null or OCarCongProc3=10 or OCarCongProc3=1291 or
OCarCongProc3=1305) then
            Return 1; -- Isolated CAB procedure.
        else
            Return null;
        end if;
    else
        -- OpValve can only be 1 at this point.
        if UnplProc=3 then
            If (VSAV=2 or VSAV is null) or (VSAV=1 and UnplAV=1) then
                if (VSMV=2 or VSMV is null) or (VSMV=1 and UnplMV=1) then
                    if (OCarCongProc1 is null or OCarCongProc1=10 or OCarCongProc1=1291 or
OCarCongProc1=1305) and
                        (OCarCongProc2 is null or OCarCongProc2=10 or OCarCongProc2=1291 or
OCarCongProc2=1305) and
                        (OCarCongProc3 is null or OCarCongProc3=10 or OCarCongProc3=1291 or
OCarCongProc3=1305) then
                        Return 1; -- Isolated CAB procedure.
                    else
                        Return null;
                    end if;
                end if;
            end if;
        end if;
    end if;
end if;

-- Procedure is not an isolated CABG, but could still be a valve or combination CAB + Valve procedure.

-- Determine whether the procedure is an isolated AVR or AVR + CAB. ProcID=2 or 4.
If OpValve=2 or OpValve is null then
    Return null; -- If procedure is not an isolated CAB and no valve procedures were done, it is an
Other procedure.
else
    if VSAV=1 and (VSAVPr=1 or VSAVPr=9) then
        if (VSMV=2 or VSMV is null) or (VSMV=1 and UnplProc=3 and UnplMV=1) then
            if (OpCAB=2 or OpCAB is null) or (OpCAB=1 and UnplProc=3 and UnplCABG=1) then
                if (OCarCongProc1 is null or OCarCongProc1=10) and (OCarCongProc2 is null or
OCarCongProc2=10) and (OCarCongProc3 is null or OCarCongProc3=10) then
                    Return 2; -- Isolated AVR procedure.
                else
```

SQL code to create function to identify procedures.txt

```
        Return null;
    end if;
else
    -- OpCAB can only be 1 at this point.
    If (Unpl Proc=3 and (Unpl CABG=2 or Unpl CABG is null)) or (Unpl Proc=1 or Unpl Proc=2 or
Unpl Proc is null) then
        if (OCarCongProc1 is null or OCarCongProc1=10 or OCarCongProc1=1291 or
OCarCongProc1=1305) and
            (OCarCongProc2 is null or OCarCongProc2=10 or OCarCongProc2=1291 or
OCarCongProc2=1305) and
            (OCarCongProc3 is null or OCarCongProc3=10 or OCarCongProc3=1291 or
OCarCongProc3=1305) then
            Return 4;    -- AVR + CAB procedure.
        else
            Return null;
        end if;
    end if;
end if;
end if;
end if;

-- Determine whether the procedure is an isolated MVR or MVR + CAB.  ProcID=3 or 5.
if VSMV=1 and (VSMVPr=2) then
    if (VSAV=2 or VSAV is null) or (VSAV=1 and Unpl Proc=3 and Unpl AV=1) then
        if (OpCAB=2 or OpCAB is null) or (OpCAB=1 and Unpl Proc=3 and Unpl CABG=1) then
            if (OCarCongProc1 is null or OCarCongProc1=10) and (OCarCongProc2 is null or
OCarCongProc2=10) and (OCarCongProc3 is null or OCarCongProc3=10) then
                Return 3;    -- Isolated MVR procedure.
            else
                Return null;
            end if;
        else
            -- OpCAB can only be 1 at this point.
            If (Unpl Proc=3 and (Unpl CABG=2 or Unpl CABG is null)) or (Unpl Proc=1 or Unpl Proc=2 or
Unpl Proc is null) then
                if (OCarCongProc1 is null or OCarCongProc1=10 or OCarCongProc1=1291 or
OCarCongProc1=1305) and
                    (OCarCongProc2 is null or OCarCongProc2=10 or OCarCongProc2=1291 or
OCarCongProc2=1305) and
                    (OCarCongProc3 is null or OCarCongProc3=10 or OCarCongProc3=1291 or
OCarCongProc3=1305) then
                        Return 5;    -- MVR + CAB procedure.
                    else
                        Return null;
                    end if;
            end if;
        end if;
    end if;
end if;
```

SQL code to create function to identify procedures.txt

```

        end if;
    end if;

    -- Determine whether the procedure is an AVR + MVR.   ProcID=6.
    if VSAV=1 and (VSAVPr=1 or VSAVPr=9) and VSMV=1 and VSMVPr=2 then
        if (OpCAB=2 or OpCAB is null) or (OpCAB=1 and UnplProc=3 and UnplCABG=1) then
            if (OCarCongProc1 is null or OCarCongProc1=10) and (OCarCongProc2 is null or OCarCongProc2=10)
and (OCarCongProc3 is null or OCarCongProc3=10) then
                Return 6;    -- AVR + MVR procedure.
            else
                Return null;
            end if;
        end if;
    end if;

    -- Determine whether the procedure is an MV Repair or MV Repair + CAB.   ProcID=7 or 8.
    if VSMV=1 and VSMVPr=1 then
        if (VSAV=2 or VSAV is null) or (VSAV=1 and UnplProc=3 and UnplAV=1) then
            if (OpCAB=2 or OpCAB is null) or (OpCAB=1 and UnplProc=3 and UnplCABG=1) then
                if (OCarCongProc1 is null or OCarCongProc1=10) and (OCarCongProc2 is null or
OCarCongProc2=10) and (OCarCongProc3 is null or OCarCongProc3=10) then
                    Return 7;    -- MV Repair procedure.
                else
                    Return null;
                end if;
            else
                -- OpCAB can only be 1 at this point.
                if (UnplProc=3 and (UnplCABG=2 or UnplCABG is null)) or (UnplProc=1 or UnplProc=2 or
UnplProc is null) then
                    if (OCarCongProc1 is null or OCarCongProc1=10 or OCarCongProc1=1291 or
OCarCongProc1=1305) and
                        (OCarCongProc2 is null or OCarCongProc2=10 or OCarCongProc2=1291 or
OCarCongProc2=1305) and
                        (OCarCongProc3 is null or OCarCongProc3=10 or OCarCongProc3=1291 or
OCarCongProc3=1305) then
                        Return 8;    -- MV Repair + CAB procedure.
                    else
                        Return null;
                    end if;
                end if;
            end if;
        end if;
    end if;

    -- If ProcID still has not been determined, then it is an Other procedure.   ProcID = null.
    return null;

```

SQL code to create function to identify procedures.txt

```
EXCEPTION  
  WHEN NO_DATA_FOUND THEN  
    NULL;  
  WHEN OTHERS THEN  
    Null;  
    RAISE;  
END getProclD;  
/
```

STS CABG Composite Score

Importance

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Opportunity for Improvement (Measure evaluation criterion 1b)

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. (This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.

The measure was calculated using STS data for patients undergoing isolated CABG in two consecutive time periods, July 2015 – June 2016 and July 2016 – June 2017.

The table below summarizes the distributions of the STS CABG composite score in the last four quarterly harvests for which the composite scores were calculated. The fall harvests cover data from July of the previous year until June of the current year. The spring harvests cover data in the previous calendar year.

Distribution of STS isolated CABG composite measure in the latest four STS harvests for which the measure was reported

Stat	STS Harvests*			
	Latest	Spring 2017	Fall 2016	Spring 2016
# Participant	945	1006	882	1026
# Operations	145815	150882	129972	149917
Mean	0.967	0.967	0.967	0.966
STD	0.00972	0.0109	0.0102	0.0104
IQR	0.0123	0.0142	0.0131	0.0134
Percentiles				
0%	0.919	0.923	0.917	0.912
10%	0.954	0.952	0.954	0.953
20%	0.959	0.958	0.960	0.958
30%	0.962	0.962	0.964	0.962
40%	0.965	0.965	0.966	0.965
50%	0.968	0.968	0.969	0.968
60%	0.970	0.971	0.971	0.970
70%	0.972	0.973	0.974	0.973
80%	0.975	0.976	0.976	0.975
90%	0.978	0.980	0.978	0.978

100%	0.985	0.989	0.986	0.986
US Geographic Region				
Midwest	267	283	261	290
Northeast	126	129	112	134
South	359	381	327	386
West	185	207	178	216
Other	8	6	4	0

* Composite measure analysis of each harvest uses the most recent one year of data until the end of last quarter. For example spring 2017 harvest uses data until December 2016.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. (This is required for endorsement maintenance. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.

This composite measure gauges the performance of the STS participant (typically a hospital, a hospital group, or a surgeon group). It is not a patient or operation level measure. Therefore at the composite score level, we do not provide data stratified by patient characteristics. Instead, we provide results stratified by participant characteristics.

Distribution of isolated CABG composite measures by regions, fall 2017 harvest, July 2016 - June 2017

Stat	Midwest	Northeast	South	West	Other*
# Participant	267	126	359	185	8
# Operations	33448	24800	63107	22601	1859
Mean	0.967	0.970	0.966	0.966	0.957
STD	0.00932	0.00848	0.0101	0.00958	0.0073
IQR	0.0109	0.0114	0.0119	0.0128	0.012
Percentiles					
0%	0.919	0.945	0.931	0.937	0.945
10%	0.957	0.958	0.951	0.953	0.948
20%	0.961	0.963	0.958	0.958	0.950
30%	0.964	0.967	0.962	0.961	0.952
40%	0.966	0.970	0.964	0.964	0.955
50%	0.969	0.972	0.966	0.967	0.958
60%	0.971	0.974	0.969	0.969	0.960
70%	0.973	0.976	0.971	0.972	0.962
80%	0.975	0.977	0.974	0.974	0.964
90%	0.978	0.979	0.978	0.977	0.964
100%	0.983	0.984	0.984	0.985	0.965

*Non-North American/Canadian

Distribution of isolated CABG composite measures by regions, fall 2016 harvest, July 2015- June 2016

Stat	Midwest	Northeast	South	West	Other*
# Participant	261	112	327	178	4
# Operations	32530	20875	55513	20427	627
Mean	0.968	0.971	0.966	0.965	0.960
STD	0.00911	0.00734	0.0109	0.0114	0.00938
IQR	0.0128	0.00887	0.0141	0.0144	0.00922
Percentiles					
0%	0.928	0.943	0.927	0.917	0.946
10%	0.956	0.963	0.952	0.952	0.950
20%	0.961	0.965	0.959	0.957	0.954
30%	0.964	0.968	0.962	0.961	0.958
40%	0.967	0.969	0.965	0.965	0.961
50%	0.969	0.972	0.968	0.968	0.962
60%	0.971	0.974	0.971	0.970	0.964
70%	0.974	0.975	0.973	0.972	0.965
80%	0.975	0.978	0.976	0.974	0.966
90%	0.978	0.980	0.979	0.977	0.966
100%	0.984	0.986	0.985	0.985	0.967

*Non-North American/Canadian

At the individual domain level, the risk-adjusted odds ratio associated with sex and race were:

Risk-adjusted odds ratio of mortality:

- Female (at BSA=1.8) v male (at BSA=2.0): 1.59 (95% confidence interval: 1.45-1.74)
- Black v white (including patients with race other than white, black, Asian): 1.17 (1.03 – 1.32)
- Asian v white (including patients with race other than white, black, Asian): 0.97 (0.80 – 1.19)

Risk-adjusted odds ratio of morbidity:

- Female (at BSA=1.8) v male (at BSA=2.0): 1.30 (1.24-1.36)
- Black v white (including patients with race other than white, black, Asian): 1.27 (1.18-1.36)
- Asian v white (including patients with race other than white, black, Asian): 1.16 (1.04 – 1.30)

For details of risk adjustment models, please see section 2b4.

The observed proportions of IMA use and perioperative medications were:

Observed proportions of IMA use:

- Female: 98.5% v male: 99.2%
- Black: 98.7% v non-black 99.1%:

Observed proportions of use of perioperative medications:

- Female: 92.6% v male: 92.5%
- Black: 93.2% v non-black: 92.5%

Note: Consistent with previous NQF reports, Non-North American hospitals are not included in models or percentages computations above. Results are virtually unchanged when these hospitals are included.

Quality Measurement in Adult Cardiac Surgery: Part 2—Statistical Considerations in Composite Measure Scoring and Provider Rating

Sean M. O'Brien, PhD,^a David M. Shahian, MD,^{b†} Elizabeth R. DeLong, PhD,^a Sharon-Lise T. Normand, PhD,^c Fred H. Edwards, MD,^d Victor A. Ferraris, MD,^e Constance K. Haan, MD,^d Jeffrey B. Rich, MD,^f Cynthia M. Shewan, PhD,^g Rachel S. Dokholyan, MPH,^a Richard P. Anderson, MD,^h and Eric D. Peterson, MD, MPH^a

^aDuke Clinical Research Institute, Durham, North Carolina; ^bTufts University School of Medicine, Boston, Massachusetts;

^cDepartment of Health Care Policy, Harvard Medical School, and ^dDepartment of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; ^eDivision of Cardiothoracic Surgery, University of Florida, Jacksonville, Florida; ^fDivision of Cardiovascular & Thoracic Surgery, University of Kentucky Chandler Medical Center, Lexington, Kentucky; ^gSentara Cardiovascular Research Institute, Norfolk, Virginia; ^hThe Society of Thoracic Surgeons, Chicago, Illinois; and

^hSeattle, Washington

Executive Summary

There is increasing interest among payers, patients, regulators, and providers to measure and compare cardiac surgery quality. The Society of Thoracic Surgeons (STS) Quality Measurement Task Force (QMTF) was established to develop comprehensive, summary performance measures encompassing multiple domains of quality. This report describes statistical considerations relevant to combining multiple measures into an overall composite score and then using such scores to rate providers.

The QMTF evaluated various options for combining 11 National Quality Forum (NQF)-endorsed process and outcome measures, both *within* and *across* the four domains of care chosen by the Task Force (Perioperative Medical Care, Operative Care, Risk-Adjusted Operative Mortality, and Postoperative Risk-Adjusted Major Morbidity). These methods included simple or weighted averaging, a composite opportunity model similar to that used by the Centers for Medicare & Medicaid Services (CMS), "all or none" scoring, scaled combinations, and latent variable models. Each method was illustrated using actual 2004 STS data from 133,149 coronary artery bypass procedures. Provider performance was estimated using Bayesian random-effects approaches to account for small sample size and to incorporate risk adjustment for outcomes.

Latent variable modeling failed to provide accurate estimates of provider performance when tested with actual STS data. Most other methods of combining individual measures *within* a given domain produced similar and consistent estimates of performance (Spearman rank correlations 0.95 to 0.98), and an all or none approach was selected.

Combining scores *across* domains was accomplished by rescaling and then adding the domain-specific estimates. When this methodology is applied to actual STS data, a one percentage point improvement in mortality has the same impact on the overall composite score as does an 8% improvement in the morbidity rate, an 11% improvement in the frequency of internal mammary artery usage, or a 28% change in the frequency of using all four NQF-recommended medications.

The QMTF considered various approaches to determining performance tiers based on composite scores. As a demonstration of one such system, the QMTF conducted a pilot study with 2004 STS data, using a 99% Bayesian certainty criterion to assign performance tiers. This stringent criterion was used to maximize the statistical certainty of tier assignments. Applying this methodology, approximately 77% of providers fell into a middle-performance tier, 10% were determined to be in a high-performing tier, and another 13% in a low-performing tier.

In summary, the STS QMTF has developed and tested a composite measure of cardiac surgery quality that encompasses multiple domains of care, uses Bayesian random-effects analyses, uses all or none scoring where appropriate, and avoids subjective weighting of individ-

Address correspondence to Dr Shahian, The Society of Thoracic Surgeons, 633 N Saint Clair St, Suite 2320, Chicago, IL 60611; e-mail: shahian@comcast.net.

†Dr Shahian is the Quality Measurement Task Force Chair and Writing Group Leader.

ual measures. One possible methodology for assigning performance tiers derived from these scores was demonstrated in a pilot study. This overall methodology was applied to actual STS data and appeared to satisfy multiple criteria for validity. These quality measures for cardiac surgery should prove useful to STS participants, payers, and governmental agencies.

Introduction

More than 15 years ago, the STS was one of the first specialty organizations to recognize the importance of developing a prospectively maintained clinical data registry. The resulting STS National Adult Cardiac Surgery Database (STS NCD) has achieved widespread acceptance by the provider community as well as interested third parties, including health policy researchers, government regulators, accrediting agencies, and payers.

The STS now faces a similar leadership opportunity as the American health care system embarks on an unprecedented effort to measure and improve quality. A major focus of this collective effort will be performance measurement, as emphasized by the latest Institute of Medicine quality report, *Performance Measurement: Accelerating Improvement* [1]. To meet such new challenges and opportunities, it will be necessary to develop quality measures for cardiac surgery that are far more comprehensive than simple risk-adjusted mortality. This set of individual and composite quality measures must be evidence-based, derived from state of the art analytic methods, and subjected to rigorous empirical evaluation.

In 2005, the STS commissioned a Quality Measurement Task Force (QMTF) to develop methods for combining multiple dimensions of performance into a single comprehensive summary quality measure. Part 1 of this QMTF report describes the conceptual framework within which the QMTF conducted its deliberations and the guidelines used to select a set of individual quality measures for coronary artery bypass grafting (CABG) [2]. In Part 2, the QMTF focuses on the following statistical and methodologic issues: (1) the distribution and correlation of selected performance measures in an actual STS NCD data sample; (2) alternative approaches to combining measures *within* and *across* the four selected domains of quality (including sensitivity analyses); and (3) various approaches for assigning providers to performance tiers based on their composite scores, including a practical example of one such method.

General Methodology

Performance Measure Definitions

The QMTF selected 11 individual CABG performance measures (five process and six outcome), all of which were endorsed by the National Quality Forum (NQF) and are available in the STS NCD (Table 1). Specific NQF inclusion/exclusion criteria were applied to these measures to the extent possible. The internal mammary artery (IMA) measure excluded patients undergoing re-

Table 1. Individual Measures and Domains in the STS Composite Quality Score

Operative Care Domain:

- Use of at least one internal mammary artery graft

Perioperative Medical Care Domain:

- Preoperative β -blockers
- Discharge β -blockers
- Discharge antiplatelet medication
- Discharge antilipid medication

Risk-Adjusted Mortality Domain

- Operative mortality

Risk-Adjusted Major Morbidity Domain

- Prolonged ventilation (>24 hours)
- Deep sternal wound infection
- Permanent stroke
- Renal insufficiency
- Reoperation

STS = Society of Thoracic Surgeons.

peat CABG surgery, the permanent stroke measure excluded patients with a previous cerebrovascular accident, and the three discharge medication measures were only calculated among patients who survived until discharge.

Study Population

Using an actual sample from the STS NCD, the QMTF investigated the distribution of individual performance measures, methodologies for combining measures *within* and *across* domains, and the results derived from one potential methodology for performance tier assignment. The 530 providers that performed at least 10 isolated CABG surgeries during 2004 (median, 195; mean, 251; range 11 to 1513) and had less than 5% missing data for each of the five NQF process measures constituted the data source. The term *provider* is used generically to refer to an STS database participant (the unit of analysis), which may be a hospital or a cardiac surgery group, or both. The final study population consisted of all 133,149 patients who underwent isolated CABG surgery by one of these providers during 2004.

"True" Process Compliance and Risk-Adjusted Outcomes Rates

The STS data for each NQF process measure consist of the number of patients cared for by a provider who were eligible to receive the specified care process (denominator) and the number of patients for whom the care process was actually delivered (numerator). Although the numerator and denominator are directly observable, the true quantity of interest relevant to performance evaluation is unobservable and may be regarded as the underlying true probability of delivering the care process. For each provider, the QMTF used analytic methods described subsequently that focus on estimating these unobservable parameters, which correspond to the five NQF process measures. In this report, the term *usage rate* will usually denote the actual observed percentage of eligible patients in which a provider used the care pro-

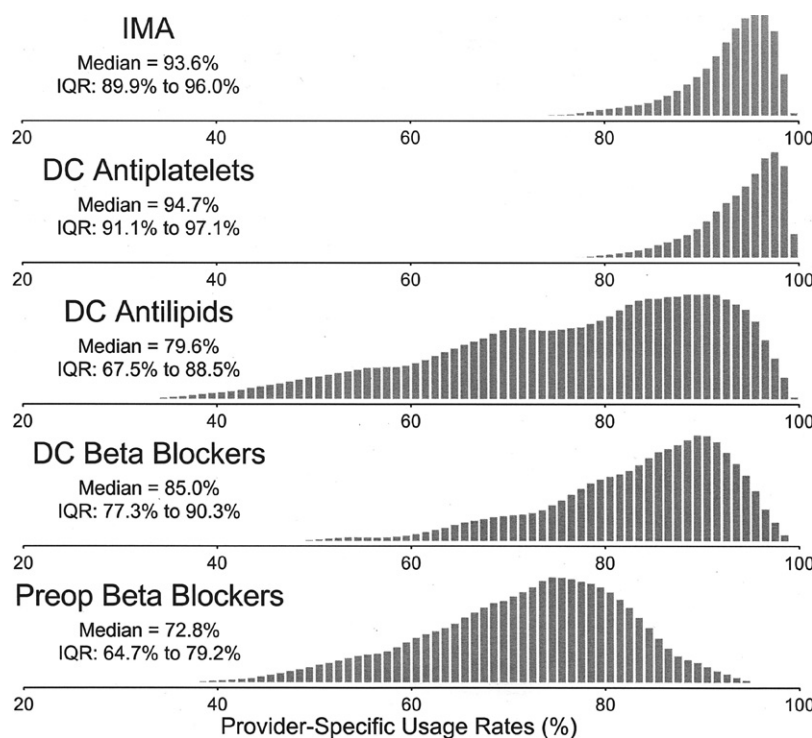


Fig 1. Estimated distribution of true provider-specific compliance rates for National Quality Forum process measures. (DC = discharge; IMA = internal mammary artery; IQR = interquartile range.)

cess, and *true usage rate* will denote the estimated corresponding “true” value.

In the case of outcomes measures, the STS data consist of the number of patients meeting the measure-specific inclusion criteria (denominator) and the number of these patients who avoided a particular adverse outcome (numerator). The “true” underlying probability of adverse outcomes may be defined and estimated in a fashion similar to that described for process measures. For outcomes, however, this estimate must also take into account the provider’s case mix, resulting in a risk-standardized adverse event rate. This may be regarded as the percentage incidence of adverse outcomes that would be anticipated if the provider treated patients having an overall risk profile similar to the STS national average. Because there are six NQF outcome measures, we define and estimate six corresponding theoretical risk-standardized rates for each provider. Estimating these parameters requires a statistical model, as described subsequently and in the Technical Appendix.

Analytical Methods

Multivariate random-effects models were applied to STS data to estimate true provider-specific usage rates for process measures and true risk-standardized event rates for outcome measures. The term *multivariate* refers to the fact that several quality measures are analyzed together in a single model, not estimated one-at-a-time in separate models. Unlike conventional methods, multivariate random-effects modeling incorporates information from

all peer providers, thereby “borrowing strength” to obtain a more reliable estimate of a single provider’s performance [3–7]. Provider-specific estimates are shrunk towards the mean for all providers, with the amount of shrinkage being inversely related to number of CABG cases and also dependent on the relative amounts of between-provider and within-provider variation.

Provider-specific risk-adjusted (or risk-standardized) mortality and morbidity rates were estimated using risk scores from previously published risk-adjustment models [8] as described in the Technical Appendix.

Bayesian methodology [9–11] was used to fit each random-effects model and to study the characteristics of alternative methods for combining and weighting quality measures. One major advantage of Bayesian approaches is that inferences about a provider’s performance are explicitly stated in terms of probabilities. For example, based on a provider’s data, we might be 99% sure that their true performance is better than average. Conventional *p* values and confidence intervals do not have a similar probability interpretation.

Distribution of Individual Performance Measures

The distribution of provider-specific performance on NQF measures was investigated by using random-effects modeling to identify overall performance levels and to quantify between-provider variation. In general, measures that vary widely yield high statistical power for discriminating among providers. Care is needed to en-

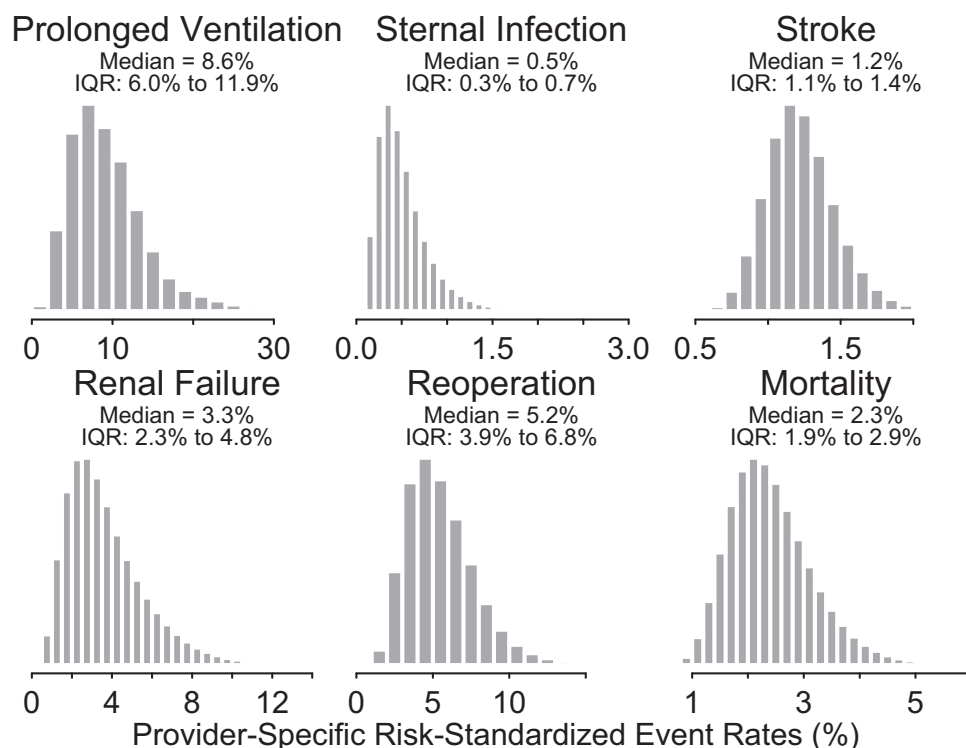


Fig 2. Estimated distribution of true provider-specific risk-standardized adverse event rates for National Quality Forum outcomes measures. (IQR = interquartile range.)

sure that the less variable measures contribute additional statistical information when they are combined with more variable measures in a composite.

The estimated distribution of true usage rates for NQF process measures is depicted in Figure 1. Between-provider variability as measured by the estimated interquartile range (IQR = 75th percentile minus 25th percentile) was greatest for the discharge antilipids measure (IQR, 67.5% to 88.5% = 21.0%), followed by preoperative β -blockers (IQR, 64.7% to 79.2% = 14.5%) and discharge β -blockers (IQR, 77.3% to 90.3% = 13.0%). The least variable measures were IMA usage (IQR, 89.9% to 96.0% = 6.1%) and discharge antiplatelets (IQR, 91.1% to 97.1% = 6.0%). Although most individual process measures had high overall estimated compliance rates, less than half of all patients received all four medications (estimated provider-specific median = 47.5%; Figure 3).

The estimated distribution of true risk-standardized event rates for NQF outcome measures is depicted in Figure 2. As previously discussed, these estimates were derived from multivariate random-effects models by using STS risk factors. For operative mortality, there is an estimated sevenfold difference in the true risk-standardized rate for the worst performing provider compared with the best (5.3% versus 0.8%; IQR 1.9% to 2.9%, median = 2.3%). The least variable outcomes measures include stroke (median, 1.2%; IQR, 1.1% to 1.4% = 0.3%) and infection (median, 0.5%; IQR, 0.3% to 0.7% = 0.4%). The most variable outcomes measure was prolonged ventilation (median, 8.6%; IQR, 6.0% to 11.9% = 5.9%).

Individual Performance Measure Correlation

The estimated correlation between pairs of NQF measures (true process usage rates and true risk-standardized outcome rates) is summarized in Tables 2 and 3. For process measures, the estimated Pearson correlation between pairs ranged from 0.10 (IMA versus discharge antiplatelets) to 0.50 (preoperative β -blockers versus discharge β -blockers). For risk-adjusted outcome measures, the estimated Pearson correlation between pairs ranged from 0.15 (prolonged ventilation versus permanent stroke) to 0.65 (sternal infection versus operative mortality).

These results suggest that individual process and outcome performance measures were generally not related to performance on other measures. Even for the most strongly correlated measures, a provider's performance on one measure did not accurately predict performance on another measure. For example, among providers who ranked in the top quartile of performance for preoperative β -blocker usage, 23.5% ranked in the bottom half of performance for discharge β -blockers. These findings suggest that the 11 selected measures provide complementary rather than redundant information about performance.

Composite Scoring Methodologies

The QMTF selected four quality-of-care domains for CABG, represented by 11 NQF-endorsed process and

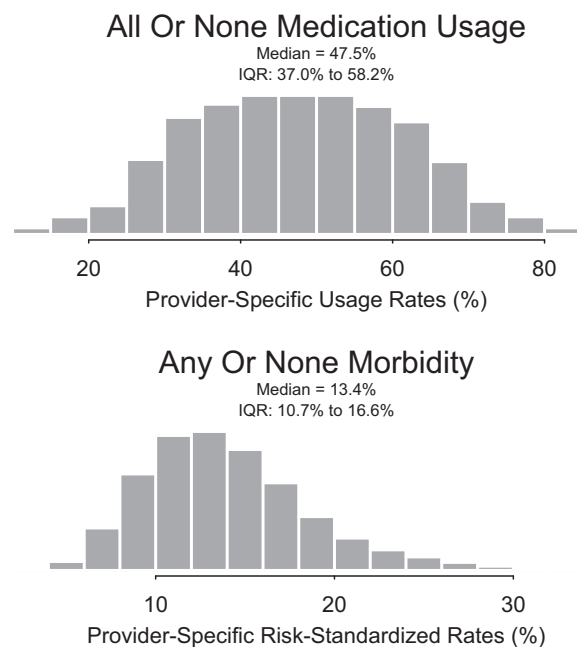


Fig 3. Estimated distribution of true provider-specific rates for all or none or any or none measures. (IQR = interquartile range.)

outcomes measures: (1) Perioperative Medical Care (pre-operative and discharge β -blockers, discharge antiplatelets, and discharge antilipids), (2) Operative Care (IMA usage), (3) Risk-Adjusted Operative Mortality, and (4) Postoperative Risk-Adjusted Major Morbidity (stroke, prolonged ventilation, renal insufficiency, reexploration for any cause, and deep sternal wound infection). The QMTF considered both (a) alternative methods to combine measures *within* multiple-measure domains (including sensitivity analyses) and (b) alternative methods to combine estimates *across* all four domains into a single composite quality score, including rescaling.

In investigating how best to determine composite scores within and across domains, the QMTF considered a variety of existing approaches from health care, educational and psychological testing, psychometrics, and public sector performance assessment [12-22], the latter predominantly from the United Kingdom and Europe.

Finally, composite score methodologies were also pilot tested using actual 2004 STS data to assess the sensitivity of rankings to the choice of methodology and to illustrate

one potential methodology for classifying providers into performance tiers.

Methods for Within-Domain Composite Scoring

For two of the four quality-of-care domains (Perioperative Medical Care and Postoperative Risk-Adjusted Major Morbidity), it was necessary to combine multiple measures into a single composite domain score. Options considered for combining individual measures *within* a domain included (a) the CMS opportunity model [23], (b) averaging of item-specific estimates, (c) all or none scoring [24], and (d) latent variable analysis [16-18, 25].

(a) CMS Opportunity Model. An *opportunity-based approach*, such as the method used by CMS in recent pay-for-performance pilot studies [23], is one way of accounting for the fact that some patients may be ineligible for some measures. An opportunity-based measure is obtained by summing the numerators for each indicator (ie, number of patients who received the particular care), summing their denominators (number of eligible patients), and dividing the former by the latter. Implicitly, each item is weighted in proportion to the percentage of eligible patients. In the case of NQF cardiac surgery measures, for which there are few recognized exclusions, almost all patients are eligible for all process measures. In this case, the opportunity-based approach should be virtually identical to simple averaging, and this was confirmed in pilot testing with actual STS data.

(b) Averaging of Item-Specific Estimates. Both simple averaging of item-specific estimates and the opportunity-based approach may permit high performance on some measures to mask poor performance on other measures that may be critical to quality (compensability). Weighting the item-specific estimates by their importance may mitigate this problem; however, the rational assignment of such weights is highly problematic. As there are no clear data available in the literature from which to derive such weights empirically, the QMTF considered alternative methods. These included an expert opinion survey of STS members conducted on behalf of the QMTF. The results of this survey demonstrated consensus among STS members (experts) on use of an IMA graft as the most important marker of CABG process quality. In contrast, there was insufficient agreement to differentiate among the other NQF process measures or the various postoperative complications. Finally, purely subjective

Table 2. Estimated Correlation Between True Success Rates for National Quality Forum Process Measures

	Pearson Correlation (95% credible interval)			
	Discharge Antiplatelets	Discharge Antilipids	Discharge β -Blockers	Preoperative β -Blockers
IMA	0.10 (0.06, 0.15)	0.15 (0.11, 0.19)	0.13 (0.09, 0.17)	0.24 (0.02, 0.28)
Discharge antiplatelets		0.38 (0.35, 0.42)	0.30 (0.27, 0.34)	0.15 (0.12, 0.19)
Discharge antilipids			0.34 (0.31, 0.37)	0.19 (0.16, 0.23)
Discharge β -blockers				0.50 (0.47, 0.53)

IMA = internal mammary artery.

Table 3. Estimated Correlation Between True Success Rates for National Quality Forum Risk-Adjusted Outcome Measures

	Pearson Correlation (95% Credible Interval)				
	Sternal Infection	Stroke	Renal Failure	Reoperation	Mortality
Prolonged ventilation	0.46 (0.29, 0.64)	0.15 (−0.14, 0.44)	0.49 (0.41, 0.58)	0.49 (0.41, 0.56)	0.50 (0.37, 0.63)
Sternal infection		0.16 (−0.25, 0.58)	0.16 (−0.02, 0.36)	0.55 (0.36, 0.75)	0.65 (0.41, 0.84)
Stroke			0.40 (0.08, 0.69)	0.43 (0.12, 0.7)	0.43 (0.06, 0.74)
Renal failure				0.44 (0.34, 0.54)	0.54 (0.39, 0.68)
Reoperation					0.61 (0.46, 0.75)

assignment of weights to individual items by the QMTF was considered but rejected as scientifically indefensible. In the absence of a clear rationale for assigning weights, experts generally consider equal weighting as the most appropriate default approach [22].

As a satisfactory weighting methodology was unavailable, averaging with equal weights was further investigated with sensitivity analyses, using the 2004 STS NCD data. A provider's overall score for medications was defined as the average of the provider's four medication-specific usage probabilities, as estimated with multivariate random-effects modeling. Similarly, a provider's overall score for the risk-adjusted morbidity domain was defined as the average of the provider's five risk-standardized event probabilities, as estimated with multivariate random-effects modeling. As summarized in columns A and B of Table 4, the results of these analyses were generally consistent with those derived from the CMS and all or none methodologies. Together with the theoretical objections as noted, this led the QMTF to reject the use of both simple and weighted averaging.

(c) All or None Scoring. The QMTF also considered the use of all or none scoring as advocated by the Institute for Healthcare Improvement [24] and the Institute of Medicine [1]. With an all or none score, performance on process measures is defined by the percentage of patients who received *all* of the care items for which they were

eligible. An analogous measure for outcomes is *any* or none, defined by the percent of patients who were discharged without having sustained *any* of the five major complications. No partial credit is given if a patient experiences some but not all of the desired results.

Application of this approach to actual STS data revealed that there was inter-provider variability of the all or none compliance percentages for the perioperative process domain and of the any or none occurrence percentages for the morbidity domain (Fig 3). This variability suggests that such a scoring approach may be useful in helping to distinguish performance differences among providers. The all or none approach yielded similar composite scores to those obtained from averaging or from an opportunity model (Table 4; Columns A and B; Spearman correlation, 0.95 to 0.98).

(d) Latent Trait Analysis, including Item Response Theory. Finally, the QMTF also considered more complex modeling techniques originally developed in the fields of psychometric and educational testing, including latent trait analysis and item response theory [16–18, 25]. Latent trait analysis is theoretically well suited for the study of an abstract construct such as aptitude or quality. In this approach, multiple observable indicators such as process compliance or morbidity rates are assumed to be related to an underlying (unobserved) latent variable such as surgical quality, the latter being the primary focus of

Table 4. Sensitivity Analyses

	A Perioperative Medical Care Domain All or None Versus CMS or Simple Average	B Postoperative Morbidity Domain Any or None Versus Simple Average	C Overall Composite Score Rescale by SD Versus Range	D Overall Composite Score Rescale by SD Versus No Rescaling
Spearman rank correlation	0.98	0.95	0.99	0.84
Rank changes by				
>50 places	8.7%	24.9%	2.1%	50.6%
>100 places	0.8%	4.5%	0.0%	24.5%
>200 places	0.0%	0.0%	0.0%	3.2%
Top 1/3 by one, not other	10.2%	11.9%	5.7%	25.0%
Top 1/3 by one, bottom 1/3 by other	0.0%	0.0%	0.0%	2.3%

CMS = Centers for Medicare & Medicaid Services; SD = standard deviation.

interest. This type of model potentially allows quality to be estimated with high statistical efficiency by combining information from multiple observable measures into a single parameter. The relative weights for each observable indicator are determined iteratively from the model, obviating the need to make a priori weight assignments.

Although latent trait modeling has potential statistical and practical advantages for discriminating among providers, the underlying assumptions (eg, unidimensionality, local independence) may not be appropriate for all data sets and must be tested on a case-by-case basis. Using 2004 STS data, we fitted the latent-trait logistic model described by Landrum and associates [17] and others. Informal model assessment included graphing the observed versus predicted values, whereas formal model evaluation consisted of computing the difference between observed and expected rates and an approximate Bayesian posterior p value.

Separate analyses were conducted for the four NQF medication measures and five NQF risk-adjusted morbidity end points. In both cases, there were large discrepancies between the model-based estimates and each provider's actual observed data, and the adequacy of each model was rejected with high confidence (Bayesian $p < 0.00001$). In contrast, the multivariate random-effects model appeared to fit adequately (Bayesian $p = 0.43$ for analysis of medications and $p = 0.38$ for morbidities). These findings suggest that one or more of the major assumptions (eg, unidimensionality, local independence) underlying the latent trait logistic model may not be tenable for STS data. This led the QMTF to reject the latent variable modeling approach for STS composite quality measures.

Final Within-Domain Composite Scoring Method

After testing all four potential methods for combining measures within domains, the QMTF selected an all/any or none approach. This method is straightforward and intuitive, avoids subjective weighting, sets an appropriately high benchmark for the ideal CABG hospitalization, and performs as well as or better than other methods when applied to actual STS data. A provider's score for the perioperative medication domain is its estimated true probability of delivering all four NQF medications. The provider's score for the morbidity domain is its estimated true risk-standardized probability of avoiding all five major morbidities.

Determination of Final Composite (Across-Domain) Scores

The next step of this project was to combine the two process measures (IMA usage rate and all or none medication compliance rate) and two risk-standardized outcomes measures (operative mortality rate and any or none morbidity rate) into a single comprehensive quality score. To assure consistent directionality, so that increasingly positive values reflect better performance, mortality rates were converted to survival rates (risk-standardized survival rate = $100 - \text{risk-standardized mortality rate}$), and morbidity rates were converted to "absence of mor-

bidity" rates (risk-standardized absence of morbidity rate = $100 - \text{risk-standardized morbidity rate}$). A provider's score for the mortality domain is the provider's risk-standardized survival rate. Similarly, the provider's score for the morbidity domain is the risk-standardized absence of morbidity rate.

Another major statistical consideration was how to account for the differing scales of measurement of the domain-specific scores. In theory, each measure has the same scale, which ranges from 0% to 100%; however, in reality, measurement scales differ dramatically. Medication adherence rates are widely dispersed and range from close to 0% to almost 100% (a range of 100%). In contrast, risk-standardized survival rates are tightly clustered in a narrow interval ranging from about 95% to 99% (a range of 4%). To account for these differences, the scales of measurement need to be standardized before the domain scores are combined into an overall composite score.

The QMTF considered multiple options to standardize measurement scales among domains and to create a single overall composite score. In the CMS approach (*a*), separate process and outcomes composite scores are averaged together, with each domain composite weighted according to the number of items it encompasses:

(*a*) The rescaled score for the j th domain is calculated as:

$$\text{RESCALED SCORE}_j = \frac{n_j}{n_1 + n_2 + n_3 + n_4} \times X_j$$

where X_j is the original score in the j th domain and n_j is the number of items comprising the j th domain. Thus,

RESCALED SCORE

$$= \frac{(\text{NUMBER OF ITEMS IN DOMAIN})}{(\text{TOTAL NUMBER OF ITEMS})} \times (\text{ORIGINAL SCORE})$$

In the case of CABG surgery, CMS uses five process measures and three outcome measures. A single summary composite is obtained by weighting the process composite by 5/8 and the outcome composite by 3/8. The QMTF regards this weighting mechanism as a significant limitation of the CMS approach. It does not account for the fact that process measure adherence and risk-standardized survival rates are measured on unequal scales. Using this approach, the outcome component contribution to the overall composite score may be substantially underweighted compared with that of the process component, which is not the desired effect. The QMTF rejected this approach.

The next two approaches involved rescaling each domain score by the reciprocal of its standard deviation (*b*) or its range (*c*), then weighting the rescaled estimates equally:

(b) Divide by the domain-specific standard deviation:

$$\text{RESCALED SCORE}_j = \frac{X_j - \bar{X}_j}{\sigma_j}$$

where \bar{X}_j is the average value of score j among STS participants and σ_j is the corresponding standard deviation. The rescaled domain-specific scores all have the same standard deviation. Thus:

$$\text{RESCALED SCORE} = \frac{(\text{ORIGINAL SCORE} - \text{AVERAGE SCORE})}{\text{STANDARD DEVIATION}}$$

(c) Divide by the domain-specific range.

$$\text{RESCALED SCORE}_j = \frac{(X_j - \text{MIN}_j)}{(\text{MAX}_j - \text{MIN}_j)}$$

where MAX_j and MIN_j are the maximum and minimum observed values of X_j across all of the providers. The rescaled domain-specific scores all lie in the interval 0 to 1. Thus:

$$\text{RESCALED SCORE} = \frac{(\text{ORIGINAL SCORE} - \text{MINIMUM})}{(\text{MAXIMUM} - \text{MINIMUM})}$$

Results were similar when rescaling was accomplished using the reciprocal of the item's range instead of its standard deviation (Spearman correlation = 0.99; column C). Ranks agree to within 100 places for 100% of providers.

(d) Results diverged to a much greater extent when the items were combined without rescaling (Spearman correlation = 0.84; column D):

$$\text{RESCALED SCORE}_j = X_j = \text{ORIGINAL SCORE}$$

Approximately 3.2% of providers changed ranks more than 200 places, depending on whether scaling was used. Furthermore, among the 177 providers that were ranked in the bottom third when rescaling was based on the standard deviation, four (2.3%) were ranked in the top third when the items were combined without rescaling.

The implications of rescaling were further explored. If items are not rescaled, then items that vary widely between providers will disproportionately influence the overall composite. Furthermore, without rescaling, a one percentage point difference in the risk-standardized mortality rate is considered to have the same importance as a one percentage point difference in the frequency of using IMA or compliance with the all or none medication measure. Rescaling by either the reciprocal of the standard deviation or the range changes the amount that improvement on a single item contributes to the overall composite. The approximate standard deviations corresponding to mortality, morbidity, IMA, and medications are 0.5, 4.2, 5.8, and 14.3, respectively. When items are weighted by the reciprocal of the standard deviation, a one percentage point improvement in mortality has the same impact on the composite score as does an 8% improvement in the morbidity rate, an 11% improvement

in the frequency of IMA usage, or a 28% change in the frequency of using all four medications. These findings are largely consistent with the QMTF's clinical assessment regarding the importance of individual items, as well as results of the STS member survey.

Final Overall Composite Scoring Method

To compute an overall composite score, the QMTF chose to rescale the domain-specific estimates by the reciprocals of their standard deviations, then add these rescaled estimates. To verify that each item contributes statistical information but does not dominate the composite, we calculated the [item-total] correlation between each domain-specific estimate and the overall comprehensive score. The [item-total] Pearson correlations were 0.48 (IMA score versus overall score), 0.56 (medication domain score versus overall score), 0.65 (morbidity domain score versus overall score), and 0.78 (mortality domain score versus overall score). Thus, although risk-adjusted mortality and morbidity explain much of the variation in the overall comprehensive score, no single item dominates, and all four items contribute statistical information.

Performance Tier Determination

Having selected a methodology for composite scoring, the QMTF considered various options for assigning providers to performance tiers. For example, a number of different approaches have been used in the United Kingdom, including confidence intervals to determine high and low outliers, ranks based on percentiles, and absolute thresholds, which, unlike the first two options, is not a reflection of performance *relative* to other providers [19, 20].

Using actual STS data, the QMTF pilot tested the discriminating power of one hypothetical three-tiered rating system. A high level of statistical certainty was deemed essential for a system designed specifically to rate providers. Accordingly, in the pilot study, providers were assigned to the middle tier if their score was statistically indistinguishable from the STS national average based on a 99% Bayesian certainty criterion. Otherwise, providers were assigned to the top tier (above average performance) or bottom tier (below average performance).

The number of providers assigned to the bottom, middle, and top tiers using this particular rating system with the 2004 STS data was 70 (13%), 407 (77%), and 53 (10%), respectively. Providers in the middle tier may be interpreted as having average performance. Their estimated performance was either very close to the STS national average value, or else the number of patients was too small to make a reliable determination. For the remaining 123 providers, the classification of above average or below average performance could be made with high confidence (more than 99% certainty). Compared with the bottom tier, providers in the top tier had lower estimated risk-standardized mortality rates (median, 1.7% versus 3.0%); lower estimated any or none morbid-

ity rates (median, 9.8% versus 18.1%); higher IMA usage rates (median, 95.7% versus 88.1%); and higher all or none medication rates (median, 66.4% versus 35.7%).

We speculated that the comprehensive quality score would have greater statistical power for discriminating between providers than a report card based solely on risk-adjusted operative mortality. The results of this pilot study confirm the statistical advantages of combining process and outcome measures when making such inter-provider comparisons. If tier assignments were to be based solely on risk-standardized operative mortality, then only 6 providers (1%) could be assigned to the top or bottom tier with at least 99% certainty. In contrast, the composite score was able to distinguish above average or below average performance for 123 providers (23%). The composite score achieves high statistical power because it combines information from 11 different quality indicators into a single estimate.

Comment

Composite indicators are useful for summarizing and comparing the quality of care delivered by healthcare providers. In many areas of medicine, the number of acceptable quality indicators is large, and there is a need for summary measures that combine performance on multiple end points. Although quality improvement requires attention to each individual aspect of quality, there are many settings in which the users of quality measures are most interested in the bottom line. The comprehensive composite quality score developed by QMTF satisfies the need for a composite quality measure for CABG providers, and it can easily be extended to valvular surgery.

Although composite measures have many practical advantages, combining multiple measures into a single indicator is inherently problematic [15, 19, 20, 22]. Similar to the development of any statistical model, considerable judgment must be exercised in the construction of a composite indicator. The choice of the individual component measures, weighting of measures, the method used to aggregate measures, and the assignment of ranks are just a few of the potential sources of controversy. Decisions regarding these and other issues may substantially impact the evaluation of providers and may have important policy implications [15, 19, 20, 22]. Furthermore, each individual indicator reflects a different aspect of quality. Therefore, some information is invariably lost when only a single summary score is reported, particularly when measures are poorly correlated and nonredundant, as they were in our pilot studies.

Depending on how the composite is constructed, some important areas of performance may not be addressed or may be relatively undervalued, and aggregation may also obscure individual areas of strength or weakness. The ability to decompose the composite into its individual components is critical. This allows providers to analyze their performance in specific areas and to formulate improvement strategies.

Finally, recent studies have demonstrated that com-

posite scores are associated with substantial random variation and may also be sensitive to factors such as the methods for weighting and aggregation [19, 20]. This emphasizes the need to correctly partition variability when calculating such scores, and to present them with appropriate measures of uncertainty.

To address the legitimate concern of whether a provider's composite performance score would be influenced substantially by the choice of statistical methodology, the QMTF implemented and compared a large number of approaches. In addition to the methods presented here, we also considered several variations of the most common approaches. In general, inferences about a provider's quality were robust and largely insensitive to the choice of methodology.

In constructing a composite scoring system, the goal of the QMTF was to use methodology that is scientifically rigorous and useful for third parties, as well as transparent, actionable, and acceptable to the cardiac surgery community. Although our goal was complete objectivity, we were often confronted by choices in which a correct decision could not be objectively determined. In such cases, detailed analyses were conducted to assess the empirical implications of our decision. In reporting these results, we have attempted to make the characteristics of the scoring system transparent to its users and to the surgeons who will be evaluated by it.

The QMTF approach to determining composite quality scores has several distinguishing features. First, performance on the individual component items is estimated using random-effects regression models. This approach, also known as shrinkage estimation, is particularly advantageous when some providers treat a small number of patients or the end points of interest are rare outcomes.

Second, to make inferences about each provider's quality, we have estimated performance using a Bayesian framework. Unlike conventional approaches, such as the CMS methodology, the Bayesian framework makes it possible to compute true probability intervals and other measures of uncertainty for any quantity of interest. Although we exploited the Bayesian framework for this investigation, our approach is flexible, and other statistical methods, such as empirical Bayes, might also be adopted.

Third, we used all or none composite scoring for the two domains of quality that contained multiple measures. This sets a high benchmark standard, the "ideal" CABG hospitalization.

Finally, for both theoretical and empirical reasons, we rejected all methods that used "importance" weighting of various measures when calculating domain scores. Neither literature review nor expert opinion survey provided scientifically valid, consistent weight determinations. Rather than subjectively assigning weights, we elected to use the all/any or none approach with implicit equal weighting. For the overall composite measure, weighting was accomplished through rescaling.

Limitations

NQF process measures have been extensively vetted to establish their suitability for performance measurement. However, it is important to acknowledge some limitations inherent to these measures and, in fact, to all measures of process compliance. Determining a set of valid exclusion criteria for process measures has proven challenging in many areas of medicine, and information on contraindications to various process measures is relatively limited in the STS NCD. As a result, some patients who are appropriately denied a medication owing to a contraindication will be misclassified as representing a process failure. The true overall percentage compliance with process measures will therefore be underestimated because some ineligible patients are included. Because the NQF measure set does not account for most exclusions, the QMTF acknowledges that all programs will have less than perfect process compliance scores. However, unless the proportion of patients who might satisfy legitimate exclusion criteria varies substantially among programs, comparisons that focus on *relative* rather than absolute process compliance will be unbiased. As exclusion criteria become more standardized and are added to the STS NCD, future versions of the QMTF scoring methodology may be modified to more fully account for them.

Limitations also exist with respect to various outcome measures. Because certain outcomes are difficult to define precisely, it is possible that variation in coding practices could account for some of the observed differences between providers.

Future Directions

Although the QMTF composite quality score is described as “comprehensive,” we do not presume that it captures every important aspect of quality. The goal was to combine the NQF measures in a statistically reasonable, practical, and intuitive fashion. The decision to create an overall composite score using four domains of care (two process domains and two outcomes domains) resulted from our commitment to use all the NQF-endorsed CABG measures that were captured in the STS NCD. The 11 relevant NQF measures are divided about equally into process and outcomes, and they appear to group naturally and appropriately into the domains we selected. The QMTF believes that these four domains assess different and not necessarily congruent aspects of the care process, all of which are important in a multidimensional quality construct.

There are multiple alternative approaches that could have been used to develop the composite score, or that may be the focus of subsequent research:

1. use of statistical methodologies (eg, factor analysis) to reduce the number of variables comprising the composite score;
2. construction of separate composite scores for processes and outcomes, assessing their degree of congruence, and studying each separately to

determine their ability to discriminate overall performance;

3. consideration of new or revised NQF measures, or non-NQF measures, that provide important incremental information about overall care; and
4. ongoing assessment of the relationship between process measures and short- or long-term outcomes, with potential elimination of process measures that demonstrate limited clinical effectiveness.

Having developed the initial methodology described in this report, all subsequent modifications can be implemented in response to STS interests and external requests.

As the STS NCD evolves from a registry used primarily for research and for internal quality improvement to one that is also used for reporting to the public and to specific third parties such as payers, the need for aggressive audit and validation is correspondingly increased. This is currently a major national initiative of the STS NCD.

The management of missing data merits careful further consideration. The implications of such missing data increase when multiple outcomes are measured, and the missing-at-random assumption is less tenable. The consequences of restricting our pilot study to sites with less than 5% missing data were not specifically investigated, but this will be an important consideration when actual composite scores are publicly released. STS NCD policies may be modified to maximize provider data completeness for elements that comprise the composite score, and the most appropriate statistical methodologies for managing any remaining missing data will be implemented.

It will be instructive to monitor provider-specific scores longitudinally to assess their consistency, and to study which domains and individual measures appear most predictive of subsequent provider performance.

If composite scores are to be used for a rating system, choices must be made regarding the use of an internal or external benchmark, the number of rating tiers, the cut points used to define each tier, and the statistical criteria used to assign providers to tiers. The hypothetical three-tiered rating system described in this report is only one of many such potential systems, and the operating characteristics of alternative approaches should be investigated and compared.

Finally, research is needed to identify the optimum time window for estimating and reporting performance. In our analyses, performance was estimated based on a single year of data. When estimating performance, methods that incorporate data from multiple time points deserve consideration.

Conclusion

This two-part report by the STS QMTF describes the development of a multidimensional CABG composite quality score that is scientifically rigorous, uses NQF-endorsed measures from the STS NCD, and is consistent with relevant national guidelines for performance mea-

surement. The QMTF regards this as the first step in a process that will constantly evolve as new quality measures, statistical methodologies, and health care policy objectives are developed.

References

1. Institute of Medicine. Performance measurement: accelerating improvement. Washington, DC: The National Academies Press; 2006.
2. Shahian DM, Edwards FH, Ferraris VA, et al. Quality measurement in adult cardiac surgery: Part 1—Conceptual framework and measure selection. *Ann Thorac Surg* 2007;83:S3–12.
3. Normand S-LT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc* 1997; 92:803–14.
4. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg* 2001;72:2155–68.
5. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med* 1997;127:764–8.
6. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc (series A)* 1996;159:385–443.
7. Leyland AH, Goldstein H. Multilevel modelling of health statistics. Chichester, UK: John Wiley and Sons, Ltd; 2001.
8. Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *Ann Thorac Surg* 2003;75:1856–64.
9. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. West Sussex, UK: John Wiley and Sons, Ltd; 2004.
10. Carlin BP, Louis TA. Bayes and empirical Bayes methods for data analysis. Boca Raton, FL: Chapman & Hall/CRC Press; 2000.
11. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Boca Raton: Chapman & Hall/CRC Press; 2004.
12. Lied T, Malsbary R, Eisenberg C, Ranck J. Combining HEDIS indicators: a new approach to measuring plan performance. *Health Care Financ Rev* 2002;23:117–29.
13. Zaslavsky A, Shaul J, Zaborski L, Cioffi M, Cleary P. Combining health plan performance indicators into simpler composite measures. *Health Care Financ Rev* 2002;23:101–15.
14. Bethell C, Reuland C, Halfon N, Schor E. Measuring the quality of preventive and developmental services for young children: national estimates and patterns of clinicians' performance. *Pediatrics* 2004;113:1973–83.
15. Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffman A, Giovannini E. Handbook on constructing composite indicators: methodology and user guide (OECD Statistics Working Paper). 2005. Organization for Economic Co-operation and Development (OECD) Statistics Working Paper JT00188147, STD/DOC; 2005:3.
16. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
17. Landrum MB, Bronskill SE, Normand S-LT. Analytical methods for constructing cross-sectional profiles of health care providers. *Health Serv Outcomes Res Methodol* 2000;1: 23–47.
18. Landrum MB, Normand S-LT, Rosenheck RA. Selection of related multivariate means: monitoring psychiatric care in the Department of Veterans Affairs. *J Am Stat Assoc* 2003; 98:7–16.
19. Jacobs R, Goddard M, Smith PC. Are composite measures a robust reflection of performance in the public sector? York, UK: Centre for Health Economics Working Paper 016, University of York; 2006.
20. Jacobs R, Goddard M, Smith P. How robust are hospital ranks based on composite performance measures? *Med Care* 2005;43:1177–84.
21. Snelling I. Do star ratings really reflect hospital performance? *J Health Organ Manag* 2003;17:210–23.
22. Booyen F. An overview and evaluation of composite indices of development. *Social Indicators Research* 2002; 59:115–51.
23. Available at: www.premierinc.com/all/quality/hqi/resources/september-scoring-overview-september.pdf. Accessed Jul 30, 2006.
24. Nolan T, Berwick DM. All-or-none measurement raises the bar on performance. *JAMA* 2006;295:1168–70.
25. Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling. Boca Raton, FL: Chapman & Hall/CRC; 2004.

TECHNICAL APPENDIX

The modelling technique adopted by the Quality Measurement Task Force is multivariate random-effects logistic regression. The term *multivariate* means all of the quality measures are analyzed together in a single model, not estimated one at a time in separate models. Random-effects refers to the assumption that the provider-specific parameters of interest are assumed to arise from a specified distribution defined by parameters that are also estimated in the modelling process. Throughout this appendix, the terms *provider* and *site* are used interchangeably to refer to The Society of Thoracic Surgeons (STS) database participants (ie, hospitals and cardiac surgery groups).

1. Final Model for Estimating Composite Scores

The following data elements were used to estimate each provider's final composite score:

1. *Risk-adjusted operative mortality.* The number of isolated coronary artery bypass grafting (CABG) patients (denominator) and the number of these patients who *did not* experience operative mortality. Note: Larger values of the numerator imply lower incidence of mortality.
2. *Risk-adjusted any or none morbidity.* The number of isolated CABG patients (denominator) and the number of these patients who *did not* experience any of the selected morbidity end points (numerator). Note: Larger values of the numerator imply lower incidence of morbidity.
3. *Internal mammary artery usage.* The number of isolated CABG patients who were eligible to receive an internal mammary artery (IMA) (denominator) and the number of these patients who actually received an IMA (numerator). Note: Larger values of the numerator imply more frequent IMA usage.
4. *All or none medications.* The number of isolated CABG patients who were eligible to receive at least one medication (denominator) and the number of these patients who received all of the medications for which they were eligible (numerator). Note: Larger values of the numerator imply more frequent use of all recommended medications.

To ensure consistent directionality between process measures and risk-adjusted outcomes, the numerators of the risk-adjusted outcomes measures are defined as the number of patients who avoid the adverse end point. Thus, for both process and outcomes, larger values of the numerator are favorable.

Summary Measures of Case Mix

In addition to the numerators and denominators listed, two summary measures of each site's case mix were also incorporated into the final multivariate random effects logistic model to risk-adjust the mortality and morbidity end points. The summary measures are:

- the average predicted risk of mortality assigned to patients at each site by the existing STS CABG mortality model; and
- the average predicted risk of "mortality or major morbidity" assigned to patients at each site by the existing STS CABG mortality/major morbidity model.

The development and validation of STS risk models is described in Shroyer and colleagues [8]. The STS CABG mortality model was updated in 2004 and has not been published. (It is available as an online STS risk calculator at www.sts.org.) Although the end point predicted by the STS mortality/major morbidity model is not identical to the any or none morbidity end point chosen by the QMTF, the STS model still provides a useful summary measure of case mix. Because mortality is a relatively rare end point, the risk factors that predict the combined end point of "mortality or major morbidity" are essentially identical to the risk factors that predict our any or none morbidity end point. In light of this similarity, we used the existing STS mortality/major morbidity risk model to calculate risk scores for risk-adjusting the any or none morbidity end point.

For each end point, the formula for calculating a patient's predicted risk of the end point has the form:

$$\text{Predicted Risk} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_q x_q}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_q x_q}}$$

where x_1, x_2, \dots, x_q denote patient risk factors (eg, quantitative variables such as age, and comorbidities coded as 1=present, 0=absent); and b_0, b_1, \dots, b_q denote regression coefficients (constants) that were determined previously (see Shroyer and colleagues [8]). Risk estimates were calculated for each patient and then averaged within providers to obtain the provider-specific average risk. A logit transformation was applied to the site-specific average risk estimates before including them in the model to express them on a scale that is not constrained by a maximum of 100% or a minimum of 0%. Thus, the final summary measures have the form $z = \log[p/(1 - p)]$, where p denotes the site-specific average risk of the endpoint.

1.1. Statistical Model

Let π_{mj} denote the true site-specific success probability at site j ($j = 1, 2, \dots, J$) for measure m , where $m = 1$ denotes avoidance of operative mortality; $m = 2$ denotes avoidance of all five morbidities; $m = 3$ denotes IMA usage; and $m = 4$ denotes use of all eligible medications. Let n_{mj} denote the number of patients who were eligible to be included in the denominator for measure m at site j , and let Y_{mj} denote the number of successful outcomes (numerator). Conditional on π_{mj} , the observed numerator is assumed to arise from a binomial distribution with probability parameter π_{mj} . That is: $Y_{mj} | \pi_{mj} \sim \text{binomial}(\pi_{mj}, n_{mj})$.

A probability model for all four outcomes Y_{1j}, \dots, Y_{4j} is obtained by assuming that each binomial outcome is conditionally independent given $(\pi_{1j}, \dots, \pi_{4j})$. Let $\mathbf{Y}_j = (Y_{1j}, \dots, Y_{4j})$ denote the collection of outcomes for the j th participant. The likelihood for site j is given by

$$\Pr[\mathbf{Y}_j = \mathbf{y}_j | (\pi_{1j}, \dots, \pi_{4j})] = \prod_{m=1}^4 \binom{n_{mj}}{y_{mj}} \pi_{mj}^{y_{mj}} (1 - \pi_{mj})^{n_{mj} - y_{mj}} \quad (1)$$

where $\mathbf{y}_j = (y_{1j}, \dots, y_{4j})$. Thus, conditional on the provider's true probability parameters $(\pi_{1j}, \dots, \pi_{4j})$, the observed data are assumed to consist of four independent binomially distributed random variables. The assumption of conditional independence is likely to be violated in practice but is made to simplify the computation. Although the model assumes *conditional* independence between (Y_{1j}, \dots, Y_{4j}) , the model does not assume *marginal* independence between these variables, because the underlying binomial parameters $(\pi_{1j}, \dots, \pi_{4j})$ are assumed to arise from a random distribution with parameters that allow for intra-item correlation.

To express probabilities on a linear scale that is not constrained by a maximum of 100% or a minimum of 0%, the probability parameters are converted to odds parameters, and we model the logarithm of the odds. Let $\theta_{mj} = \pi_{mj}/(1 - \pi_{mj})$ denote the odds of success for measure m at site j . The odds is interpreted as the probability of a successful outcome divided by the probability of a unsuccessful outcome. If the odds can be estimated, then it can be converted to a probability, because $\pi_{mj} = \theta_{mj}/(1 + \theta_{mj})$. Similar to π_{mj} , larger values of θ_{mj} imply a higher probability of a successful outcome. We focus on the logarithm of the odds parameters, $\log \theta_{mj}$, because this quantity ranges from negative infinity to infinity (ie, no boundary constraints).

A fundamental assumption of the multivariate hierarchical logistic model is that the parameters $(\log \theta_{1j}, \log \theta_{2j}, \log \theta_{3j}, \log \theta_{4j})$ are distributed according to a multivariate normal distribution. Correlation among performance on different end points is reflected in the covariance parameters of the multivariate normal distribution. We further assume that a provider's performance on any single end point is described by a logistic regression model. The latter assumption is embodied by the set of equations:

$$\begin{aligned} (\text{mortality}) \quad \log \theta_{1j} &= \alpha_1 + \beta_1 z_{1j} + \varepsilon_{1j} \\ (\text{morbidity}) \quad \log \theta_{2j} &= \alpha_2 + \beta_2 z_{2j} + \varepsilon_{2j} \\ (\text{IMA}) \quad \log \theta_{3j} &= \alpha_3 + \varepsilon_{3j} \\ (\text{medications}) \quad \log \theta_{4j} &= \alpha_4 + \varepsilon_{4j} \end{aligned}$$

where $(\alpha_1, \alpha_2, \alpha_3, \text{ and } \alpha_4)$ denote intercept parameters that determine the overall frequency of success for the four measures; $(\varepsilon_{1j}, \varepsilon_{2j}, \varepsilon_{3j}, \text{ and } \varepsilon_{4j})$ are normally distributed error terms that determine the extent to which the j th site deviates from the average; z_{1j} denotes the logit of the average predicted risk of mortality at site j , as determined by the STS mortality model (described above); z_{2j} denotes the logit of the average predicted risk of "mortality or major morbidity", as determined by the STS composite end point model (described above); and $(\beta_1 \text{ and } \beta_2)$ denote regression coefficients to be estimated from the data. The terms $\beta_1 z_{1j}$ and $\beta_2 z_{2j}$ are included to incorporate risk-adjustment into the analysis of the mortality and morbidity end points. No assumptions are made about the covariance parameters of the multivariate normal distribution. An equivalent specification of the model is:

$$\begin{pmatrix} \log \theta_{j1} \\ \log \theta_{j2} \\ \log \theta_{j3} \\ \log \theta_{j4} \end{pmatrix} \stackrel{\text{indep}}{\sim} N \left[\begin{pmatrix} \alpha_1 + \beta_1 z_{1j} \\ \alpha_2 + \beta_2 z_{2j} \\ \alpha_3 \\ \alpha_4 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix} \right]$$

where $(\sigma_{11}, \sigma_{12}, \dots, \sigma_{44})$ denote unknown parameters of the multivariate normal covariance matrix. These unknown covariance parameters are estimated from the data along with the unknown α s and β s.

1.2. Definition of Risk-Standardized Rates

In the case of risk-adjusted outcomes measures, π_{mj} is not a meaningful reflection of a site's quality because it partly reflects the site's case mix. Interest instead focuses on estimating each provider's "risk-standardized success rate," denoted by π'_{mj} . Because there is no single widely accepted definition of the risk-standardized success rate, the QMTF considered several options, and chose the following:

$$\text{Risk-standardized success rate} = \pi'_{mj} = \frac{e^{\alpha_m + \beta_m \bar{z}_m + e_{mj}}}{1 + e^{\alpha_m + \beta_m \bar{z}_m + e_{mj}}} \quad (2)$$

where \bar{z}_m denotes the value of z_{mj} for an "average" provider. The risk-standardized success rate π'_{mj} is loosely interpreted as the success rate for measure m that would be projected to occur hypothetically if provider j had a "typical" case mix.

1.3. Definition of True Composite Score

The final composite score was defined as the quantity

$$\text{Composite Score}_j = \frac{\pi'_{1j}}{c_1} + \frac{\pi'_{2j}}{c_2} + \frac{\pi'_{3j}}{c_3} + \frac{\pi'_{4j}}{c_4}$$

where $c_1 = 0.5$, $c_2 = 4.2$, $c_3 = 5.8$, and $c_4 = 14.3$. These constants were chosen such that c_m is approximately equal to the standard deviation (across providers) of the corresponding parameter, π'_{mj} or π_{mj} . Larger values of the composite score imply better performance.

1.4. Method of Estimation

The quantities π'_{1j} , π'_{2j} , π'_{3j} , π'_{4j} were estimated in a Bayesian framework by specifying a diffuse normal prior for α_1 and α_2 ; an informative normal prior distribution for β_1 , β_2 , β_3 , and β_4 ; and a diffuse Wishart prior for the distribution of $T = \Sigma^{-1}$, where $\Sigma = (\sigma_{11}, \sigma_{12}, \dots, \sigma_{44})$ denotes the covariance matrix of the random effects distribution. Specifically:

$$\alpha_m \stackrel{\text{indep}}{\sim} N(0.0, 10000.0)$$

$$\beta_m \stackrel{\text{indep}}{\sim} N(1.0, 1.0)$$

$$T \sim \text{Wishart}_4 \begin{pmatrix} 0.100, & 0.005, & 0.005, & 0.005 \\ 0.005, & 0.100, & 0.005, & 0.005 \\ 0.005, & 0.005, & 0.100, & 0.005 \\ 0.005, & 0.005, & 0.005, & 0.100 \end{pmatrix}$$

where $N(a, b)$ denotes a normal distribution with mean a and variance b ; and $\text{Wishart}_\nu(R)$ denotes a Wishart distribution with ν degrees of freedom and scale matrix R . The Wishart distribution is parameterized such that $\Sigma_i = \frac{1}{\nu} z_i z_i' \sim \text{Wishart}_\nu(R)$ if $z_i \stackrel{iid}{\sim}$

$N(0, R)$, with R denoting the covariance matrix of the multivariate normal distribution of the z_i .

The chosen scale matrix implies that the prior mean of the correlation between two random effects from the same site is equal to 0.05, that is, $E[\text{corr}(\epsilon_{mj}, \epsilon_{m'j})] = 0.05$. According to the prior distribution, there is also 50% prior probability that $\text{corr}(\epsilon_{mj}, \epsilon_{m'j})$ lies in the interval $(-0.70, 0.70)$. The parameters $\{\alpha_m\}$, $\{\beta_m\}$, and T were assumed to be mutually independent in the prior distribution.

The $N(1,1)$ prior for β_m is motivated by the fact that $\beta_m = 1$ by definition under the assumption that each provider's true event rate is exactly equal to the rate predicted by the STS risk model; hence, we chose our prior mean to be 1.0. In reality, we do not believe $\beta_m = 1$. (Owing to site-level variation in performance, we do not believe the STS risk model will exactly predict each site's true event rate). The prior variance of 1.0 was chosen to allow for uncertainty regarding the true value of β_m . Although larger values of the variance might be considered desirable (because larger variance implies greater uncertainty about β_m), we encountered computational difficulties (slow mixing) with larger variance.

2. Models for Combining Items Within a Domain (Reported but not Selected)

Although the QMTF ultimately chose to combine items within a domain by using the all or none method, a variety of other model-based approaches were considered. These included: (i) fitting a multivariate random effects model and then averaging the item-specific estimates; and (ii) fitting a latent trait logistic model similar to the one described by Landrum and colleagues [17]. Each of these two modeling strategies was applied separately to the perioperative medication domain (four items analyzed simultaneously: preoperative β -blockers, discharge β -blockers, discharge antiplatelets and discharge antilipids) and the major morbidity domain (five items analyzed simultaneously: prolonged ventilation, sternal infection, stroke, renal insufficiency, and reoperation). Altogether we fit four models (2 types of models \times 2 domains).

To describe these models, let M denote the total number of measures considered in a single model ($M = 4$ for medication models; $M = 5$ for morbidity models); and let J ($= 530$) denote the number of providers. Using the notation of Part 1, let n_{mj} , y_{mj} , π_{mj} , and θ_{mj} denote the number of eligible patients (denominator), the number of successful results (numerator), the true success probability, and the odds of success, respectively, for measure m ($m = 1, 2, \dots, M$) and site j ($j = 1, 2, \dots, J$). The probability model for site j 's data is given by equation (1) above. Each of the four models described below was estimated in a Bayesian framework using vague proper priors for the distribution of model parameters.

MODEL A1. MULTIVARIATE RANDOM EFFECTS MODEL FOR ANALYZING THE FOUR NQF MEDICATION MEASURES.

($M = 4$.) Model:

$$\begin{aligned} (\text{preop } \beta\text{-blockers}) \quad & \log \theta_{1j} = \alpha_1 + \epsilon_{1j} \\ (\text{discharge } \beta\text{-blockers}) \quad & \log \theta_{2j} = \alpha_2 + \epsilon_{2j} \\ (\text{discharge antiplatelets}) \quad & \log \theta_{3j} = \alpha_3 + \epsilon_{3j} \\ (\text{discharge antilipids}) \quad & \log \theta_{4j} = \alpha_4 + \epsilon_{4j} \end{aligned}$$

where $(\epsilon_{1j}, \dots, \epsilon_{4j})$ are distributed according to a multivariate normal distribution having mean vector zero and an unstructured covariance matrix.

MODEL A2. MULTIVARIATE RANDOM EFFECTS MODEL FOR ANALYZING THE FIVE NQF MORBIDITY MEASURES.

($M = 5$.) Model:

$$\begin{aligned} (\text{prolonged ventilation}) \quad & \log \theta_{1j} = \alpha_1 + \beta_1 z_{1j} + \varepsilon_{1j} \\ (\text{infection}) \quad & \log \theta_{2j} = \alpha_2 + \beta_2 z_{2j} + \varepsilon_{2j} \\ (\text{stroke}) \quad & \log \theta_{3j} = \alpha_3 + \beta_3 z_{3j} + \varepsilon_{3j} \\ (\text{renal failure}) \quad & \log \theta_{4j} = \alpha_4 + \beta_4 z_{4j} + \varepsilon_{4j} \\ (\text{reoperation}) \quad & \log \theta_{5j} = \alpha_5 + \beta_5 z_{5j} + \varepsilon_{5j} \end{aligned}$$

where z_{mj} denotes a summary measure of site j 's case mix used for risk-adjusting measure m (defined below); and $(\varepsilon_{1j}, \dots, \varepsilon_{5j})$ are distributed according to a multivariate normal distribution with mean vector zero and an unstructured covariance matrix.

The summary measures of case mix (z_{1j}, \dots, z_{5j}) were calculated from previously validated STS risk models, using the approach described in Section 1 of the Appendix. Separate STS risk models exist for each of the five NQF-endorsed morbidity end points (Shroyer and colleagues [8]). The quantity z_{mj} is defined as the logit of the average predicted risk of end point m at site j .

MODEL A3. LATENT TRAIT LOGISTIC MODEL FOR ANALYZING THE FOUR NQF MEDICATION MEASURES.

($M = 4$.) Model:

$$\begin{aligned} (\text{preop } \beta\text{-blockers}) \quad & \log \theta_{1j} = \alpha_1 + \gamma_1 Q_j \\ (\text{discharge } \beta\text{-blockers}) \quad & \log \theta_{2j} = \alpha_2 + \gamma_2 Q_j \\ (\text{discharge antiplatelets}) \quad & \log \theta_{3j} = \alpha_3 + \gamma_3 Q_j \\ (\text{discharge antilipids}) \quad & \log \theta_{4j} = \alpha_4 + \gamma_4 Q_j \end{aligned}$$

where each Q_j is independently distributed according to a normal distribution with mean zero and unit variance, ie, $Q_j \stackrel{iid}{\sim} N(0, 1)$; and for identifiability we assume that $\gamma_1 > 1$. This model is an application of the latent trait logistic model described by Landrum and colleagues [17]. In this model, Q_j represents the "latent quality" of the j th site.

MODEL A4. LATENT TRAIT LOGISTIC MODEL FOR ANALYZING THE FIVE NQF MORBIDITY MEASURES.

($M = 5$.) Model:

$$\begin{aligned} (\text{prolonged ventilation}) \quad & \log \theta_{1j} = \alpha_1 + \beta_1 z_{1j} + \gamma_1 Q_j \\ (\text{infection}) \quad & \log \theta_{2j} = \alpha_2 + \beta_2 z_{2j} + \gamma_2 Q_j \\ (\text{stroke}) \quad & \log \theta_{3j} = \alpha_3 + \beta_3 z_{3j} + \gamma_3 Q_j \\ (\text{renal failure}) \quad & \log \theta_{4j} = \alpha_4 + \beta_4 z_{4j} + \gamma_4 Q_j \\ (\text{reoperation}) \quad & \log \theta_{5j} = \alpha_5 + \beta_5 z_{5j} + \gamma_5 Q_j \end{aligned}$$

where z_{mj} denotes a summary measure of site j 's case mix used for risk-adjusting measure m (defined above); each Q_j is independently distributed according to a normal distribution with mean zero and unit variance, ie, $Q_j \stackrel{iid}{\sim} N(0, 1)$; and for identifiability, we assume that $\gamma_1 > 1$. This model is a slight generalization of the latent trait logistic model described by Landrum and colleagues [17]. The model they described did not include the terms $\beta_m z_{mj}$ that allow for risk adjustment.

Use of Bayesian p -Values to Test Fit of Latent Trait Logistic Models

The Bayesian p -value is the probability that a hypothetical replicated data set, y^{rep} , would diverge from the true model as much

as the observed data set, y , diverges from the true model. Divergence between the model and the data was defined by the quantity

$$D(y, \pi) \equiv \sum_{j=1}^J \sum_{m=1}^M \frac{n_{mj}(y_{mj}/n_{mj} - \pi_{mj})^2}{\pi_{mj}(1 - \pi_{mj})}$$

where π denotes the collection of all of the parameters π_{mj} . The Bayesian p -value was defined as

$$\begin{aligned} \text{Bayesian } p\text{-value} &= \Pr[D(y^{\text{rep}}, \pi) \geq D(y, \pi) | y] \\ &= \int \Pr[D(y^{\text{rep}}, \pi) \geq D(y, \pi) | \pi, y] p(\pi | y) d\pi \\ &= \int \left\{ \int I_{D(y^{\text{rep}}, \pi) \geq D(y, \pi)} p(y^{\text{rep}} | \pi) d y^{\text{rep}} \right\} p(\pi | y) d\pi \end{aligned}$$

where I is the indicator function; $p(y^{\text{rep}} | \pi)$ is the probability density function for a hypothetical replicated data set conditional on the model parameters (defined by equation 1, above); and $p(\pi | y)$ is the posterior distribution of the model parameters given the observed data. The probability of interest is taken over the joint distribution $p(y^{\text{rep}}, \pi | y)$. To calculate the probability of interest, the integral

$$\begin{aligned} \Pr[D(y^{\text{rep}}, \pi) \geq D(y, \pi) | \pi, y] \\ = \int I_{D(y^{\text{rep}}, \pi) \geq D(y, \pi)} p(y^{\text{rep}} | \pi) d y^{\text{rep}} \end{aligned} \quad (3)$$

was approximated as $\Pr[\chi_{dM \times J}^2 \geq D(y, \pi) | \pi, y]$, where χ_d^2 denotes a χ^2 random variable with d degrees of freedom. The final approximate p -value was calculated as

$$\text{Bayesian } p\text{-value} \approx \frac{1}{N} \sum_{i=1}^N \Pr[\chi_{dM \times J}^2 \geq D(y, \pi^{(i)})]$$

where $\pi^{(1)}, \pi^{(2)}, \dots, \pi^{(N)}$ denote N draws from a Markov chain Monte Carlo simulation with target distribution $p(\pi | y)$.

3. Models for Other Results Reported

Other univariate and multivariate random effects models were used to derive results reported in this article. The histograms in Figure 1 and the correlations in Table 2 were derived from a multivariate random effects model that was identical to Model A1 (described above), except that it included the IMA end point in addition to the medication measures. Figure 2 and Table 3 were derived from a multivariate random effects model that was identical to Model A2 (described above), except that it included the mortality end point in addition to the morbidity measures.

The final multivariate model for estimating composite performance (described in Part 1 of the Appendix) involves four end points: operative mortality, any or none morbidity, IMA usage, and all or none medications. In addition to analyzing these end points simultaneously in a single multivariate model, we also analyzed these end points one at a time by fitting four separate univariate random effects models. The method of incorporating risk adjustment was identical to the method described for the multivariate model. These univariate analyses were used to produce the top panel of Figure 3 (all or none medication usage); the bottom panel of Figure 3 (any or none morbidity end point); and to count the number of sites that would be identified as outliers if performance was estimated based on operative mortality alone (last paragraph of Performance Tier Determination section).

The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: Part 1—Coronary Artery Bypass Grafting Surgery

David M. Shahian, MD,^a Sean M. O'Brien, PhD,^b Giovanni Filardo, PhD, MPH,^c Victor A. Ferraris, MD,^d Constance K. Haan, MD,^e Jeffrey B. Rich, MD,^f Sharon-Lise T. Normand, PhD,^g Elizabeth R. DeLong, PhD,^b Cynthia M. Shewan, PhD,^h Rachel S. Dokholyan, MPH,^b Eric D. Peterson, MD, MPH,^b Fred H. Edwards, MD,^e and Richard P. Anderson, MD^{i†}

^aMassachusetts General Hospital, Boston, Massachusetts; ^bDuke Clinical Research Institute, Durham, North Carolina; ^cInstitute for Health Care Research and Improvement, Baylor Health Care System, Dallas, Texas; ^dUniversity of Kentucky Chandler Medical Center, Division of Cardiovascular and Thoracic Surgery, Lexington, Kentucky; ^eUniversity of Florida, Division of Cardiothoracic Surgery, Jacksonville, Florida; ^fSentara Cardiovascular Research Institute, Norfolk, Virginia; ^gDepartment of Health Care Policy, Harvard Medical School, and Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts; ^hThe Society of Thoracic Surgeons, Chicago, Illinois; and ⁱSeattle, Washington

Background. The first version of The Society of Thoracic Surgeons National Adult Cardiac Surgery Database (STS NCD) was developed nearly 2 decades ago. Since its inception, the number of participants has grown dramatically, patient acuity has increased, and overall outcomes have consistently improved. To adjust for these and other changes, all STS risk models have undergone periodic revisions. This report provides a detailed description of the 2008 STS risk model for coronary artery bypass grafting surgery (CABG).

Methods. The study population consisted of 774,881 isolated CABG procedures performed on adult patients aged 20 to 100 years between January 1, 2002, and December 31, 2006, at 819 STS NCD participating centers. This cohort was randomly divided into a 60% training (development) sample and a 40% test (validation) sample. The development sample was used to identify predictor variables and estimate model coefficients. The validation sample was used to assess model calibration and discrimination. Model outcomes included operative mortality, renal failure, stroke, reoperation for any cause, prolonged ventilation, deep sternal wound infection, composite major morbidity or mortality, prolonged length of stay (> 14 days), and short length of stay (< 6 days and alive). Candidate predictor variables were selected based on their availability in versions 2.35, 2.41, and 2.52.1 of the STS NCD and their presence in (or ability to be mapped to) version 2.61. Potential predictor

variables were screened for overall prevalence in the study population, missing data frequency, coding concerns, bivariate relationships with outcomes, and their presence in previous STS or other CABG risk models. Supervised backwards selection was then performed with input from an expert panel of cardiac surgeons and biostatisticians. After successfully validating the fit of the models, the development and validation samples were subsequently combined, and the final regression coefficients were estimated using the overall combined (development plus validation) sample.

Results. The c-index for the mortality model was 0.812, and the c-indices for other endpoints ranged from 0.653 for reoperation to 0.793 for renal failure in the validation sample. Plots of observed versus predicted event rates revealed acceptable calibration in the overall population and in numerous subgroups. When patients were grouped into categories of predicted risk, the absolute difference between the observed and expected event rates was less than 1.5% for each endpoint. The final model intercept and coefficients are provided.

Conclusions. New STS risk models have been developed for CABG mortality and eight other endpoints. Detailed descriptions of model development and testing are provided, together with the final algorithm. Overall model performance is excellent.

(Ann Thorac Surg 2009;88:S2–22)

© 2009 by The Society of Thoracic Surgeons

In 1986, The Society of Thoracic Surgeons (STS) convened an Ad Hoc Committee on Risk Factors for Coronary Artery Bypass Graft Surgery (CABG) [1] and an

Ad Hoc Committee to Develop a National Database for Cardiothoracic Surgery [2]. This was prompted by the

†This author is deceased. Former Chair, Quality, Research and Patient Safety Council, The Society of Thoracic Surgeons, Chicago, IL.

Address correspondence to Dr Shahian, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114; e-mail: dshahian@partners.org.

Drs Shahian, O'Brien, Filardo, Ferraris, Haan, Rich, Normand, DeLong, Shewan, Peterson, Edwards, Anderson, and Ms Dokholyan have no conflicts of interest to declare regarding this work.

Abbreviations and Acronyms

BSA	=	body surface area
CABG	=	coronary artery bypass graft surgery
CHF	=	congestive heart failure
EF	=	ejection fraction
GFR	=	glomerular filtration rate
HCFA	=	Health Care Financing Administration
IABP	=	intra-aortic balloon pump
NYHA	=	New York Heart Association
NCD	=	National Adult Cardiac Surgery Database
O/E	=	observed to expected ratio
QMTF	=	Quality Measurement Task Force
STS	=	The Society of Thoracic Surgeons

release earlier that year of inadequately risk-adjusted hospital mortality data by the Health Care Financing Administration (HCFA), now the Centers for Medicare and Medicaid Services. Although the HCFA analytical methodology was widely criticized, STS leadership recognized that the underlying principle of collecting and analyzing data to improve patient outcomes was valid, particularly for complex and costly procedures such as coronary artery bypass grafting surgery (CABG). They believed that it was the responsibility of professional organizations to develop credible clinical data registries for their own specialties, and that risk models derived from such registries would circumvent many of the concerns resulting from the use of unadjusted administrative data. Such clinical registries would be used as credible data sources for quality assessment and improvement activities as well as for research.

These early activities ultimately led to the development of the STS National Adult Cardiac Surgery Database (NCD) [3, 4]. Since its release to members in 1990, the STS NCD has evolved to become one of the largest specialty-specific clinical data registries in the world. It currently has more than 950 participants enrolled, representing just under 90% of the cardiac surgery providers in the United States, with data on more than 3.6 million procedures. Similar STS data registries have now been developed for congenital heart surgery and general thoracic surgery, and future plans include the development of specialty modules (eg, quality metrics, atrial fibrillation surgery, thoracic aortic surgery). Recent enhancements, including the addition of unique physician and patient identifiers, will facilitate linkages with other registries and greatly expand the potential of the STS NCD for longitudinal follow-up, comparative effectiveness, and cost efficiency studies.

In addition to the development of the STS NCD as a comprehensive, nationally representative data registry, the second major goal of the STS was to assure that analyses derived from this registry would be appropriately adjusted for preoperative patient severity, a major deficiency of the HCFA reports that were initially published in 1986. This was accomplished by first identifying

risk factors for specific procedures and outcomes, beginning with isolated CABG, then using these predictor variables to develop risk models. With statistical risk models, which are most often based on logistic regression, the expected outcome for a patient with a given set of risk factors can be determined, and that can be compared with the observed outcome. The observed (O) and expected (E) outcomes are summed over all patients of a particular surgeon or hospital to yield the risk-standardized mortality ratio (O/E), which can then be multiplied by the average rate in the reference population to calculate risk-standardized mortality rates [5–7].

STS CABG risk models have undergone periodic updates and revisions, the most recent of which was based upon 2000 to 2002 STS NCD data. In 2007, the STS Database Modernization Task Force completed a major specification upgrade of the STS NCD data collection instrument from version 2.52.1 to version 2.61. This included refinement, modification, consolidation, or elimination of some data elements, as well as an attempt to harmonize definitions with those of the American College of Cardiology National Cardiovascular Data Registry whenever possible. Given these changes, as well as the number of years since the last risk model update, the STS Quality Measurement Task Force (QMTF) was asked to develop new risk models for isolated CABG, isolated valve repair or replacement, and combined CABG plus valve procedures. The authors of this report include the QMTF members who participated in this initiative.

Implementation of these new models in January 2008 coincided with the release of STS NCD version 2.61. This report, Part 1 of 3, describes the development of the new mortality and morbidity models for isolated CABG surgery.

Study Purpose

The primary goal of this study was to develop risk-prediction algorithms for patients undergoing isolated CABG surgery. As the major intended use of these algorithms was to compare participant outcomes to the overall STS national experience, risk factors were generally restricted to patient and clinical characteristics present preoperatively.

Risk Model Development and Transparency

The availability of user-friendly statistical software programs and the exponential increase in computing speed have greatly facilitated statistical analyses such as logistic regression, the basis for many risk models. However, despite these technological advances, clinical judgment, experience, intuition, and practicality still play a critical role in risk model development. There are many points in model development at which legitimate differences in approach may lead to substantial differences in the resulting statistical models and the inferences derived from them [8].

We believe the degree of transparency provided in this report regarding the development of the STS CABG risk

models is essential in today's health care environment. In an era when society demands full transparency regarding health care performance, the methodologies used to evaluate that performance should be just as transparent [9, 10]. This fundamental principle is among the standards established by the American Heart Association and American College of Cardiology for statistical models used for public reporting [11].

Study Population and Endpoints

All isolated CABG procedures performed on adult patients aged 20 to 100 years between January 1, 2000, and December 31, 2006, were initially considered for inclusion, although the final development and validation samples were derived from 2002 to 2006 data. Patients missing data on sex ($n = 195$) were excluded, as these patients are not included in STS performance feedback reports to database participants. That left a study population of 774,881 surgical procedures from 819 database participants. Patients on dialysis preoperatively ($n = 12,415$) were excluded when developing the risk model for postoperative renal failure.

Training and Validation Samples

The study population was randomly divided into a 60% training (development) sample and a 40% test (validation) sample. The development sample was used to identify predictor variables and estimate model coefficients. Data from the validation sample were used to assess model fit, discrimination, and calibration. After choosing variables and assessing model fit, the development and validation samples were subsequently combined, and the final model coefficients were estimated using the combined (development plus validation) data.

Endpoints

Risk models were developed for the nine endpoints listed below. Only mortality was recorded beyond the index hospitalization. Morbidity data included only in-hospital complications, although beginning in STS NCD version 2.61, sternal infections will be recorded for up to 30 days postoperatively. The nine endpoints are as follows: (1) operative mortality: death during the same hospitalization as surgery, regardless of timing, or within 30 days of surgery regardless of venue; (2) permanent stroke (cerebrovascular accident): a central neurologic deficit persisting longer than 72 hours; (3) renal failure: a new requirement for dialysis or an increase of the serum creatinine to more than 2.0 mg/dL and double the most recent preoperative creatinine level; (4) prolonged ventilation (longer than 24 hours); (5) deep sternal wound infection; (6) reoperation for any reason; (7) major morbidity or mortality: a composite defined as the occurrence of any of the above endpoints; (8) prolonged postoperative length of stay (PLOS): length of stay (LOS) more than 14 days (alive or dead); and (9) short postoperative LOS (SLOS): LOS less than 6 days and patient alive at discharge (this SLOS definition differs from the previous STS risk models, which did not exclude patients who died in-hospital; patients who died within 5 days of surgery are

included in the new models but are treated as not having a short stay).

Table 1 summarizes the frequencies of these endpoints in the study population for each predictor variable category (ie, the bivariate relationships).

Selection of Candidate Predictor Variables

Initial Data Screening of Candidate Predictor Variables

We began by considering all possible candidate variables from the development set (Table 2). Because the primary goal of the STS risk models is to adjust surgical outcomes, in general only preoperative patient variables are included. However, because these models are also used for other purposes such as individual patient prediction and counseling, there were a few modifications (which are discussed in the relevant sections) in the application of this general principle.

As there were a large number of procedures and endpoints available, we were not statistically constrained to highly parsimonious models, nor is such an approach generally favored in regression modeling [12–14]. Discarding valid data elements can waste valuable information that has been collected at substantial effort and cost. Furthermore, although much of the discrimination of a predictive model may be contained in a relatively small number of variables [15, 16], some predictor variables that add only modestly to discrimination may still be important predictors of outcomes at the patient level [17, 18].

Expert Panel Review for Clinical Relevance and Face Validity

All candidate variables available in version 2.61 were individually discussed by a panel of cardiac surgeons and health policy experts to assure that clinical relevance as well as multiple aspects of validity (face, construct, and content) had been considered.

Data Version for Model Development

Although these new risk models were to be introduced in conjunction with the release of STS NCD version 2.61, they were developed with data collected under the three previous data versions (2.35, 2.41, and 2.52.1) because no 2.61 data were yet available. The QMTF began its predictor selection process with two caveats. First, any candidate variable had to be collected consistently across the three previous data versions. Second, it had to also be available in version 2.61 or have the ability to be mapped to this new version. For example, history of smoking and renal failure were not candidate variables as they were either not included in, or were unable to be mapped to, version 2.61. Renal function is now assessed by the last preoperative serum creatinine value, which is collected in all data versions. Because the definition of hypercholesterolemia has changed substantially over successive STS data versions, and because counterintuitive results have been observed in some previous analyses of hyper-

Table 1. Distribution of Risk Factors and Frequency of Adverse Outcomes in Overall Study Population, Isolated Coronary Artery Bypass Graft Surgery (2002–2006)

Variable	Number of Patients		Percent of Patients Experiencing Endpoint								
	N	%	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
Overall											
Total	774,881	100.0	2.3	1.4	3.6	9.7	0.4	5.2	14.4	5.6	51.2
Age, years											
< 55	137,318	17.72	1.0	0.5	1.7	7.1	0.3	3.7	10.0	2.7	67.1
55–64	221,697	28.61	1.3	0.9	2.4	7.8	0.4	4.2	11.4	3.8	59.4
65–74	245,132	31.63	2.4	1.6	3.9	10.0	0.5	5.5	14.9	5.9	47.7
≥ 75	170,734	22.03	4.7	2.3	6.4	13.9	0.5	7.5	20.9	9.6	33.0
Sex											
Male	560,006	72.27	2.0	1.2	3.4	8.7	0.4	5.1	13.4	4.9	55.0
Female	214,875	27.73	3.4	1.9	4.1	12.2	0.5	5.6	17.0	7.2	41.5
Race											
Caucasian	665,941	85.94	2.3	1.3	3.5	9.3	0.4	5.1	13.9	5.3	52.2
Black	44,405	5.73	2.7	2.0	5.2	13.5	0.7	6.3	19.0	8.2	41.3
Hispanic	25,103	3.24	2.6	1.5	4.3	11.3	0.5	5.6	16.1	6.1	48.4
Asian	12,509	1.61	2.7	1.9	3.8	12.6	0.3	7.2	17.4	7.0	45.2
Other	21,222	2.74	2.3	1.3	3.6	10.4	0.5	5.5	14.8	6.0	48.7
Missing	5,701	0.74	2.3	1.4	4.1	9.4	0.4	5.2	14.5	6.1	48.9
Body surface area (m ²)											
< 1.50	14,339	1.85	6.2	2.4	4.6	16.2	0.3	8.3	22.1	9.8	36.5
1.50–1.74	111,458	14.38	3.8	2.0	4.0	12.6	0.3	6.5	17.7	7.4	42.5
1.75–1.99	280,677	36.22	2.4	1.5	3.5	9.6	0.4	5.4	14.4	5.6	50.7
≥ 2.00	363,817	46.95	1.7	1.0	3.6	8.6	0.5	4.6	13.1	4.8	55.0
Missing	4,590	0.59	3.7	1.4	4.0	7.6	0.3	4.7	13.9	6.8	46.0
Body mass index (kg/m ²)											
< 25	169,091	21.82	3.3	1.7	3.5	11.0	0.3	6.7	16.3	6.7	47.6
25–29	303,371	39.15	2.1	1.3	3.1	8.6	0.3	4.9	13.1	4.8	54.2
30–34	186,148	24.02	1.8	1.2	3.6	9.0	0.5	4.5	13.4	5.0	53.1
≥ 35	110,213	14.22	2.3	1.2	5.2	12.0	0.8	4.9	16.8	6.8	45.7
Missing	6,058	0.78	3.7	1.4	4.2	8.6	0.3	4.8	14.5	6.7	47.2
Diabetes mellitus											
No diabetes	492,800	63.60	2.1	1.2	2.8	8.8	0.3	5.0	13.0	4.7	54.8
Diabetes–noninsulin	195,421	25.22	2.3	1.6	4.3	10.1	0.5	5.2	15.2	6.0	48.2
Diabetes–insulin	84,406	10.89	3.6	1.8	7.1	13.9	1.0	6.5	20.6	9.7	37.5
Diabetes–missing treatment	1,439	0.19	3.1	2.2	4.6	11.1	0.7	4.6	15.7	8.8	41.9
Missing	815	0.11	3.8	0.7	2.6	8.8	0.5	4.5	12.3	6.9	43.7
Hypertension											
No	167,260	21.59	1.9	0.9	2.2	8.1	0.3	4.6	11.7	4.2	58.2
Yes	606,813	78.31	2.5	1.5	4.0	10.1	0.5	5.4	15.1	5.9	49.3
Missing	808	0.10	3.8	0.7	2.4	9.3	0.5	5.2	12.7	6.7	43.9
Hypercholesterolemia											
No	199,894	25.80	3.0	1.6	3.9	11.0	0.5	5.8	16.1	6.5	48.7
Yes	573,257	73.98	2.1	1.3	3.5	9.2	0.4	5.0	13.8	5.2	52.1
Missing	1,730	0.22	4.1	1.6	3.5	10.3	0.3	4.7	13.9	7.3	47.5
Past or present smoker											
No	295,999	38.20	2.4	1.4	3.7	9.0	0.4	5.1	13.9	5.3	50.1
Yes	477,911	61.68	2.3	1.3	3.6	10.1	0.5	5.3	14.7	5.7	51.9
Missing	971	0.13	3.4	0.7	3.1	9.9	0.4	5.3	13.5	9.1	41.0
Chronic lung disease											
None	612,211	79.01	2.0	1.3	3.3	8.4	0.3	4.9	13.0	4.7	53.7
Mild	85,005	10.97	2.8	1.5	4.2	12.0	0.6	5.8	16.9	7.0	45.7
Moderate	47,745	6.16	3.8	1.6	5.3	15.8	0.8	6.8	20.8	9.6	39.5
Severe	22,302	2.88	7.0	2.0	7.7	22.8	1.1	9.5	29.0	15.3	29.2
Missing	7,618	0.98	2.6	1.4	3.2	8.2	0.2	3.9	12.6	5.7	53.4

Table 1. Continued

Variable	Number of Patients		Percent of Patients Experiencing Endpoint								
	N	%	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
Peripheral vascular disease											
No	653,260	84.30	2.0	1.2	3.2	8.8	0.4	4.8	13.1	4.8	53.5
Yes	120,480	15.55	4.4	2.3	6.1	14.4	0.7	7.5	21.2	9.6	38.7
Missing	1,141	0.15	3.9	1.1	3.1	11.8	0.3	5.5	14.4	7.4	43.9
Cerebrovascular disease											
No	668,073	86.22	2.1	1.1	3.3	9.0	0.4	4.9	13.4	5.0	53.4
Yes	105,792	13.65	4.0	2.9	5.8	14.0	0.6	7.2	20.7	9.3	37.7
Missing	1,016	0.13	3.2	0.6	2.4	8.9	0.3	4.2	11.4	6.8	43.2
CVA											
No CVA	717,721	92.62	2.2	1.2	3.4	9.3	0.4	5.1	13.8	5.2	52.5
Remote CVA (> 2 weeks)	53,341	6.88	4.2	3.1	6.1	15.3	0.7	7.4	22.0	10.3	35.5
Recent CVA (≤ 2 weeks)	1,763	0.23	5.0	4.9	6.5	18.8	0.9	8.7	25.4	13.0	32.5
CVA—missing timing	745	0.10	3.8	3.5	6.9	14.0	0.5	5.9	21.7	10.7	34.2
Missing	1,311	0.17	3.3	1.0	2.6	7.4	0.2	4.3	11.4	5.9	47.8
Endocarditis											
No endocarditis	773,002	99.76	2.3	1.4	3.6	9.7	0.4	5.2	14.4	5.5	51.2
Treated endocarditis	472	0.06	4.4	0.8	5.3	15.3	0.6	8.5	19.9	8.9	33.7
Active endocarditis	110	0.01	2.7	1.8	6.3	20.0	1.8	11.8	24.5	20.0	41.8
Endocarditis—missing type	90	0.01	4.4	4.4	5.8	11.1	1.1	3.3	15.6	2.2	55.6
Missing	1,207	0.16	4.1	1.0	3.5	9.0	0.2	4.6	12.8	7.0	46.2
Renal failure											
No	731,626	94.42	2.1	1.3	3.2	9.0	0.4	5.0	13.4	5.0	52.8
Yes	42,153	5.44	7.2	2.7	14.7	22.5	1.0	9.9	31.9	15.8	23.4
Missing	1,102	0.14	3.1	0.8	2.8	7.4	0.3	3.4	10.8	6.4	46.6
Renal function											
Creatinine < 1.00 mg/dL	274,197	35.39	1.6	1.1	1.5	8.0	0.3	4.4	11.2	4.0	55.6
Creatinine 1–1.49 mg/dL	398,833	51.47	2.0	1.3	3.4	8.9	0.4	5.0	13.5	4.9	53.1
Creatinine 1.5–1.99 mg/dL	57,779	7.46	4.5	2.3	10.8	16.1	0.7	7.8	25.2	10.6	34.5
Creatinine 2.0–2.49 mg/dL	12,463	1.61	6.9	2.9	14.3	21.3	0.9	9.4	31.5	15.3	24.7
Creatinine ≥ 2.5 mg/dL	7,906	1.02	8.2	3.2	20.4	23.4	0.9	11.1	37.9	18.6	20.4
Dialysis	12,415	1.60	8.4	2.7	*NA	25.3	1.2	10.5	31.5	16.4	19.6
Missing	11,288	1.46	3.3	1.2	3.1	7.6	0.3	4.3	12.9	5.9	50.1
Immunosuppressive treatment											
No	758,368	97.87	2.3	1.4	3.6	9.6	0.4	5.2	14.2	5.4	51.5
Yes	14,976	1.93	5.4	1.8	6.3	15.6	0.8	8.7	22.5	10.8	37.0
Missing	1,537	0.20	3.3	0.8	2.8	6.5	0.4	4.6	11.4	6.1	46.8
Prior CABG Surgery											
No	735,033	94.86	2.2	1.4	3.5	9.4	0.4	5.1	14.1	5.4	51.7
Yes	36,693	4.74	5.3	1.6	5.8	14.7	0.5	7.5	20.9	7.8	42.6
Missing	3,155	0.41	2.7	1.1	3.3	8.9	0.5	4.8	12.9	6.8	48.9
Prior valve surgery											
No	769,434	99.30	2.3	1.4	3.6	9.7	0.4	5.2	14.4	5.5	51.3
Yes	2,280	0.29	5.9	1.9	6.8	15.3	0.7	8.6	22.5	11.1	32.0
Missing	3,167	0.41	2.9	1.2	3.5	8.6	0.5	4.4	12.7	6.3	50.0
Prior other cardiac surgery											
No	755,653	97.52	2.3	1.4	3.6	9.6	0.4	5.2	14.3	5.5	51.3
Yes	15,218	1.96	3.9	1.5	4.9	13.1	0.6	6.6	18.6	7.6	45.5
Missing	4,010	0.52	2.8	1.0	2.9	8.5	0.4	4.4	12.2	5.9	50.5
Number of previous CV surgeries											
No previous CV surgery	723,623	93.39	2.2	1.4	3.5	9.4	0.4	5.1	14.0	5.4	51.7
One prior CV surgery	40,474	5.22	4.7	1.6	5.4	13.8	0.5	7.3	19.9	7.7	44.1
Two or more prior CV Surgeries	4,840	0.62	6.2	1.4	5.6	14.7	0.6	8.0	22.0	8.4	41.5
Missing	5,944	0.77	2.9	1.1	3.2	9.8	0.6	5.5	14.3	6.1	51.2

Table 1. Continued

Variable	Number of Patients		Percent of Patients Experiencing Endpoint								
	N	%	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
Prior PCI											
No PCI	606,824	78.31	2.3	1.4	3.6	9.5	0.4	5.1	14.2	5.6	51.1
PCI ≤ 6 hours	7,373	0.95	8.9	2.1	7.4	25.8	0.6	10.5	32.6	11.3	35.5
PCI > 6 hours	155,161	20.02	2.3	1.2	3.5	9.6	0.4	5.4	14.3	5.1	52.4
PCI-missing timing	2,456	0.32	3.0	0.6	3.3	8.0	0.8	5.1	13.4	6.8	48.9
Missing	3,067	0.40	3.3	1.0	3.0	9.8	0.5	4.7	13.7	6.2	47.3
Acuity status											
Elective	381,116	49.18	1.5	1.1	2.9	6.6	0.4	4.3	11.1	4.1	55.6
Urgent	356,287	45.98	2.4	1.5	3.9	10.8	0.5	5.6	15.7	6.2	48.5
Emergent	34,513	4.45	8.1	2.6	8.3	29.6	0.7	10.4	34.1	13.3	33.2
Emergent salvage	1,967	0.25	38.6	4.9	17.4	52.7	0.7	18.4	70.0	23.5	12.6
Missing	998	0.13	3.2	1.3	3.6	9.2	0.5	4.8	13.7	6.5	43.6
MI											
No prior MI	424,599	54.80	1.5	1.1	2.8	6.9	0.3	4.5	11.3	4.1	55.4
MI > 21 days	137,522	17.75	2.1	1.3	3.5	8.9	0.5	5.2	13.9	5.4	50.4
MI 8–21 days	26,205	3.38	4.0	1.8	6.0	14.4	0.8	7.5	20.7	10.2	38.1
MI 1–7 days	148,659	19.18	3.4	1.8	4.8	14.0	0.5	6.1	18.9	7.5	45.7
MI > 6 and < 24 hours	21,044	2.72	6.0	2.4	6.7	23.6	0.5	8.1	28.1	10.4	39.1
MI ≤ 6 hours	11,539	1.49	10.4	2.6	8.6	31.2	0.6	10.6	36.8	13.3	33.5
MI-missing timing	4,064	0.52	3.6	1.6	4.6	11.3	0.5	5.6	17.5	7.2	43.9
Missing	1,249	0.16	2.1	1.1	2.4	6.6	0.2	3.8	10.2	6.7	49.4
Angina											
No	130,143	16.80	2.5	1.4	3.8	9.5	0.4	5.6	14.7	6.2	48.5
Yes	643,815	83.09	2.3	1.3	3.6	9.7	0.4	5.2	14.3	5.4	51.8
Missing	923	0.12	2.3	1.0	2.6	8.8	0.5	4.0	11.2	8.2	43.9
Cardiogenic shock											
No	758,766	97.92	2.0	1.3	3.4	8.9	0.4	5.0	13.6	5.2	51.9
Yes	14,919	1.93	18.0	3.6	14.6	49.6	1.0	15.3	55.7	23.1	18.3
Missing	1,196	0.15	2.7	1.1	3.2	8.0	0.4	4.4	12.0	7.4	44.7
Resuscitation											
No	766,674	98.94	2.2	1.3	3.5	9.4	0.4	5.1	14.1	5.4	51.5
Yes	6,939	0.90	17.1	3.0	11.4	37.5	0.9	14.0	46.1	18.2	24.3
Missing	1,268	0.16	2.2	0.8	3.4	8.0	0.6	3.9	11.5	7.3	45.0
Arrhythmia											
No arrhythmia	706,709	91.20	2.0	1.3	3.3	8.9	0.4	4.9	13.4	5.0	53.1
AFib/flutter	39,125	5.05	5.4	2.3	7.1	16.4	0.7	8.5	23.8	11.9	29.4
Heart block	10,026	1.29	5.8	1.9	6.8	16.8	0.6	9.2	24.2	9.4	36.4
Sustained VT/VF	14,336	1.85	8.2	2.0	6.8	23.8	0.6	11.1	31.5	12.0	33.0
Arrhythmia–other	1,853	0.24	3.8	1.6	5.3	12.9	0.6	6.5	19.1	7.3	39.2
Arrhythmia–missing type	1,344	0.17	3.9	1.7	4.4	11.9	0.7	7.3	17.5	8.3	37.9
Missing	1,488	0.19	2.7	1.0	3.0	7.1	0.5	3.8	11.1	6.5	45.5
Preoperative IABP											
No	714,824	92.25	2.0	1.3	3.3	8.0	0.4	4.9	12.8	4.9	52.8
Yes	58,134	7.50	6.9	2.2	7.7	30.8	0.6	9.6	34.4	12.9	32.0
Missing	1,923	0.25	4.2	1.7	4.3	10.9	0.6	5.8	16.0	7.2	45.7
NYHA class											
I	97,812	12.62	1.5	1.1	2.5	6.3	0.3	4.4	10.6	3.9	57.0
II	187,947	24.25	1.3	1.1	2.6	6.5	0.3	4.2	10.7	3.8	56.5
III	287,760	37.14	2.0	1.3	3.6	9.0	0.4	5.0	13.9	5.4	51.3
IV	165,325	21.34	4.5	1.9	5.5	16.6	0.6	7.1	21.9	8.8	42.5
Missing	36,037	4.65	2.4	1.2	3.6	8.9	0.3	5.3	13.6	5.8	47.7
Congestive heart failure											
No	666,592	86.03	1.8	1.2	2.9	7.9	0.3	4.7	12.2	4.3	54.7
Yes	106,700	13.77	5.9	2.4	8.5	21.0	0.9	8.7	28.0	13.2	29.5
Missing	1,589	0.21	3.3	1.4	3.1	9.2	0.4	4.5	13.0	7.4	49.6

Table 1. Continued

Variable	Number of Patients		Percent of Patients Experiencing Endpoint								
	N	%	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
Number of diseased coronary vessels											
None	2,012	0.26	2.3	0.4	2.8	8.9	0.4	4.6	12.6	5.5	53.1
One	32,311	4.17	1.5	0.6	1.9	6.1	0.2	4.4	9.8	3.1	66.3
Two	150,881	19.47	1.8	1.0	2.7	8.0	0.4	4.5	12.0	4.5	56.3
Three	586,658	75.71	2.5	1.5	4.0	10.4	0.4	5.5	15.3	6.0	49.1
Missing	3,019	0.39	2.6	0.6	1.8	5.5	0.2	4.5	10.8	5.9	54.2
Left main disease $\geq 50\%$											
No	554,355	71.54	2.1	1.3	3.4	8.8	0.4	5.0	13.5	5.1	52.7
Yes	217,548	28.08	3.0	1.5	4.3	11.9	0.5	5.9	16.8	6.6	47.6
Missing	2,978	0.38	2.3	1.4	2.7	6.3	0.3	5.5	11.9	5.9	45.4
Ejection fraction (%)											
< 25	25,323	3.27	7.2	2.2	8.0	25.2	0.8	10.5	31.9	13.7	27.8
25–34	57,460	7.42	4.6	2.1	6.1	17.6	0.6	7.6	23.8	10.3	36.8
35–44	108,623	14.02	3.0	1.7	4.7	12.4	0.6	6.0	17.5	7.2	45.7
45–54	189,478	24.45	1.9	1.3	3.4	8.7	0.4	4.8	13.2	5.0	53.1
≥ 55	351,455	45.36	1.5	1.1	2.7	6.8	0.3	4.4	11.1	3.9	56.1
Missing	42,542	5.49	3.4	1.4	4.1	10.8	0.4	5.6	16.0	6.1	50.0
Mitral stenosis											
No	756,609	97.64	2.3	1.4	3.6	9.7	0.4	5.2	14.4	5.5	51.2
Yes	2,703	0.35	5.5	2.4	6.4	17.0	0.7	7.5	22.9	10.5	35.7
Missing	15,569	2.01	2.1	1.3	3.4	8.3	0.4	4.6	13.1	5.0	53.1
Aortic stenosis											
No	750,185	96.81	2.3	1.4	3.6	9.6	0.4	5.2	14.3	5.5	51.4
Yes	11,386	1.47	4.7	2.1	6.5	14.8	0.7	7.9	21.5	9.7	36.6
Missing	13,310	1.72	2.3	1.3	3.3	8.5	0.4	4.7	13.1	5.0	52.5
Tricuspid stenosis											
No	756,574	97.64	2.3	1.4	3.6	9.7	0.4	5.2	14.4	5.6	51.2
Yes	597	0.08	3.4	2.3	6.6	14.9	0.7	6.0	20.9	10.1	43.6
Missing	17,710	2.29	2.1	1.3	3.6	8.5	0.4	4.7	13.4	5.0	53.2
Pulmonic stenosis											
No	753,975	97.30	2.3	1.4	3.6	9.7	0.4	5.2	14.4	5.6	51.2
Yes	445	0.06	3.4	2.2	3.9	12.6	0.0	6.3	20.2	6.3	49.4
Missing	20,461	2.64	2.2	1.4	3.8	8.7	0.4	5.0	13.9	5.3	52.5
Mitral insufficiency											
None	622,173	80.29	2.1	1.2	3.3	8.9	0.4	4.9	13.4	5.0	53.2
Trivial	49,152	6.34	2.4	1.6	4.2	10.5	0.4	5.7	15.7	6.2	47.9
Mild	60,811	7.85	3.7	2.0	5.7	14.3	0.5	6.9	20.3	8.6	40.3
Moderate	16,723	2.16	6.7	2.7	7.9	20.1	0.7	9.6	28.0	12.5	30.3
Severe	2,143	0.28	8.7	3.1	8.9	24.1	0.6	11.0	32.6	15.1	28.2
Missing	23,879	3.08	2.1	1.2	3.0	7.5	0.4	4.7	12.0	5.2	51.5
Aortic insufficiency											
None	705,771	91.08	2.3	1.3	3.5	9.5	0.4	5.1	14.1	5.4	51.9
Trivial	17,988	2.32	3.6	2.1	5.6	13.4	0.5	7.0	19.4	8.3	40.9
Mild	18,571	2.40	4.1	2.2	5.9	14.3	0.4	7.3	20.8	9.0	37.9
Moderate	3,576	0.46	5.3	2.6	7.0	16.2	0.5	7.9	23.2	10.2	32.8
Severe	411	0.05	7.1	1.9	6.7	15.6	0.7	9.5	25.8	10.9	37.2
Missing	28,564	3.69	2.1	1.2	3.2	7.9	0.4	4.6	12.5	5.3	51.4
Tricuspid insufficiency											
None	675,778	87.21	2.2	1.3	3.4	9.4	0.4	5.1	13.9	5.3	52.1
Trivial	32,856	4.24	2.5	1.6	4.5	11.1	0.4	6.1	16.6	6.7	47.4
Mild	29,611	3.82	3.9	2.2	5.9	14.7	0.5	7.4	21.0	9.1	39.3
Moderate	5,753	0.74	7.6	3.0	9.0	22.7	0.5	9.8	30.2	13.7	26.9
Severe	728	0.09	9.1	2.9	10.5	24.2	0.4	10.9	33.0	17.2	26.2
Missing	30,155	3.89	2.2	1.2	3.2	7.9	0.4	4.6	12.6	5.2	51.7

Table 1. Continued

Variable	Number of Patients		Percent of Patients Experiencing Endpoint								
	N	%	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
Pulmonic insufficiency											
None	724,258	93.47	2.3	1.4	3.6	9.7	0.4	5.2	14.3	5.5	51.4
Trivial	10,726	1.38	2.8	1.5	4.3	12.3	0.5	6.5	17.3	7.3	44.8
Mild	4,867	0.63	3.8	2.1	5.6	14.1	0.4	7.4	21.0	9.1	39.7
Moderate	546	0.07	6.6	3.1	7.8	17.6	0.2	7.9	24.4	11.5	29.7
Severe	217	0.03	5.1	0.5	5.1	9.7	0.5	6.5	15.7	8.8	50.7
Missing	34,267	4.42	2.2	1.3	3.5	8.3	0.4	4.8	13.2	5.5	50.7

AFib = atrial fibrillation; CABG = coronary artery bypass graft surgery; Comp = composite adverse outcome (any); CV = cardiovascular; CVA = cerebrovascular accident (stroke); DSWI = deep sternal wound infection; IABP = intra-aortic balloon pump; MI = myocardial infarction; Mort = mortality; Na = not applicable; NYHA = New York Heart Association; PCI = percutaneous coronary intervention; PLOS = prolonged length of stay; Reop = reoperation; RF = renal failure; SLOS = short length of stay; Vent = prolonged ventilation; VF = ventricular fibrillation; VT = ventricular tachycardia.

cholesterolemia, a decision was made not to include this variable in the new models.

Predictor Frequency

For each variable, the QMTF explored the overall prevalence and missing data frequency per year. Predictor variables that are rarely present in the development sample are difficult to model. For this reason, mitral (0.35%), tricuspid (0.08%), and pulmonic stenosis (0.06%), pulmonic insufficiency (0.10%), and endocarditis (0.09%) were not considered as variables in the new isolated CABG models.

Inconsistently Coded Variables

A few variables have been collected inconsistently or with questionable reliability, often for clinically unavoidable reasons. For example, pulmonary artery mean pressure data were missing for 70% of patients during 2002 to 2006. Furthermore, the value of this continuous variable may vary substantially depending on the clinical state and volume-loading status of the patient when the measurement is obtained. Because of these concerns, pulmonary artery pressure was not included in the models.

Derived or Redundant Variables

Several derived variables were considered for inclusion in the models. For example, body mass index (BMI) is a useful measure of overall body habitus. However, because BMI is highly correlated with body surface area (BSA), the more commonly used anthropometric measure in most previous STS models, the latter was retained in the new models. Similarly, there is a theoretical superiority to inclusion of glomerular filtration rate (GFR) rather than serum creatinine as a measure of renal function. However, the Modification of Diet in Renal Disease formula for estimating GFR is a complex function of creatinine, race, sex, and age, and not all laboratories perform this calculation automatically. Furthermore, as age, sex, and race are already model covariates, using GFR would complicate the interpretation of their regression coefficients. Some of the prognostic value of GFR

comes from these variables that are already included in the model. Finally, previous studies suggest that various measures of renal function used in CABG mortality risk models have similar performance [19]. For all these reasons, serum creatinine was retained as the measure of renal function.

Controversial Variables

RACE. Several variables raised particular clinical, statistical, or health policy issues. For example, race was an obvious candidate variable because it was a significant predictor ($p < 0.001$) of each endpoint except mortality and because the proportion of nonwhite patients varied substantially across institutions. In exploratory analyses, the association between race and outcomes persisted after adjusting for hospital identity, suggesting that this association is not explained by differences in hospital quality.

However, general principles of risk model development complicated the decision as to whether or not to include race in the models. When the dominant purpose of a risk model is adjustment of provider results, it is advisable to include only biological and clinical patient variables that are present before a patient's first contact with the provider. In this context, race is clearly a fixed biological characteristic, but its impact on patient outcomes may be mediated through other mechanisms. It is possible that certain racial and ethnic groups have worse outcomes not because of inherent biological characteristics but because of differences in the quality of care delivered to them. In this case, including race and ethnicity in a risk model could essentially select out or obscure the very disparity issues that society wishes to identify and correct. Inclusion of race and ethnicity in a risk model would say, in effect, that we expect nonwhites to have inferior results and would make an allowance for providers who care for such patients, just as we would for providers who care for patients in cardiogenic shock.

After deliberation regarding the pros and cons, the QMTF ultimately elected to retain race and ethnicity in the new models because of their impact on outcomes,

Table 2. Initial List of Potential Candidate Variables

Demographics

1. Age
2. Sex
3. Race (black, Caucasian, Hispanic, Asian, Native American, other)

Note: Data collection changed in v2.61. New version allows for multiple races (check all that apply). Added Hawaiian/Pacific Islander category. Hispanic ethnicity is a separate variable.

Anthropometric

4. Height
5. Weight

Status

6. Status (elective, urgent, emergent, salvage)
7. Shock
8. Resuscitation

Cardiac variables

9. Angina, angina type (STS categories are unstable, stable, no angina)

Note: Angina was removed on v2.61 data collection form. The new form has a variable called "cardiac presentation on admission." Angina is one of possible response categories to that field.

10. New York Heart Association functional class

Note: In v2.61, NYHA class is only collected if patient has congestive heart failure.

11. Arrhythmia and arrhythmia type (sustained VT/VF; heart block; AFib/flutter, None)
12. Myocardial infarction timing: (≤ 6 , > 6 and < 24 hours; 1–7, 8–21, > 21 days)

Hemodynamic/catheterization variables

13. Ejection fraction
14. Number of diseased vessels (0, 1, 2, 3)
15. Left main disease
16. Pulmonary artery mean pressure
17. Mitral stenosis
18. Aortic stenosis
19. Tricuspid stenosis
20. Pulmonic stenosis
21. Mitral insufficiency (none, trivial, mild, moderate, severe)
22. Aortic insufficiency (none, trivial, mild, moderate, severe)
23. Tricuspid insufficiency (none, trivial, mild, moderate, severe)
24. Pulmonic insufficiency (none, trivial, mild, moderate, severe)

Comorbidities

25. Serum creatinine
26. Dialysis
27. Renal failure
- Note: This variable was removed in v2.61.*
28. Endocarditis (active, treated, none)
29. Diabetes and treatment (insulin, oral, diet, untreated, no diabetes)
30. Chronic lung disease (none, mild, moderate, severe)
31. Congestive heart failure
32. Peripheral vascular disease
33. Cerebrovascular disease
34. CVA and CVA timing (recent, remote, none)

Note: CVA is a child field of cerebrovascular disease in v2.61.

35. Hypercholesterolemia (v2.35, v2.41) and Dyslipidemia (v2.52)

Note: Data from all 3 versions were merged and analyzed under the variable name "hypercholesterolemia."

36. Hypertension
37. Smoker

Note: Major definition change in v2.61.

Preoperative interventions

38. Preoperative intra-aortic balloon pump
39. Preoperative inotropes
40. Immunosuppressive treatment
41. Prior percutaneous coronary intervention and timing (≤ 6 hours, > 6 hours, none)

Previous Interventions

42. Prior coronary artery bypass graft surgery
43. Prior valve surgery
44. Prior other cardiac surgery
45. Number of previous cardiovascular surgeries

while recognizing the potential limitations of this decision.

PREOPERATIVE INTRA-AORTIC BALLOON PUMP. Preoperative intra-aortic balloon pump (IABP) is a proxy for more serious preoperative status of the patient (eg, unstable angina, ventricular dysfunction). It captures information that may not be present in other data elements, and it is associated with higher risk of postoperative morbidity and mortality. For these reasons, most CABG risk models include preoperative IABP as a risk predictor. However, placement of an IABP is also a highly discretionary care process the frequency of which varies widely among participating institutions. Indications are subjective and are often dictated by the cardiologist before even referring the patient for cardiac surgery. Based on CABG risk models, an institution that liberally utilizes IABPs will have a higher expected risk of morbidity and mortality (according to the model) compared with another institution with a similar case-mix but a more restrictive IABP policy. That would impact their relative O/E ratios and risk-adjusted outcomes.

Despite its discretionary nature (and the potential for gaming), the QMTF decided to retain IABP use in the models because it is such an important predictor. Ultimately, it was elected to model preoperative IABP as a joint variable with preoperative inotrope use as an overall measure of preoperative acuity/severity.

Review of External Sources

The QMTF also reviewed multiple external resources to aid in the selection of potential candidate variables [15, 16, 20]. First, all previous versions of the STS CABG risk models were reviewed. The QMTF also examined other CABG risk models including the European System for Cardiac Operative Risk Evaluation (EuroSCORE) [21], the New York Cardiac Surgery Reporting System [22], the Veterans Affairs Administration cardiac surgery models [23, 24], and the Northern New England Cardiovascular Disease Study Group model [25, 26]. We particularly wanted to identify variables that were found in some form across all the risk models. Subject to the constraints of version 2.61 data specifications, we made a special effort to include such variables in the new STS risk models, in some instances requiring us to “force” them into the models, as described in the section on the final variable selection procedure.

Missing Data

Missing data in the STS NCD are rare, having a frequency of less than 1% for most variables. Candidate predictor variables missing most commonly were ejection fraction (5.5%), New York Heart Association (NYHA) class (4.7%), tricuspid insufficiency (3.9%), aortic insufficiency (3.7%), mitral insufficiency (3.1%), aortic stenosis (1.7%), and creatinine/dialysis (1.5%).

Missing predictor values in the STS NCD were managed using imputation. Multiple imputation is the generally preferred statistical method [27], but single imputation was also considered based on the following

practical considerations: (a) the fraction of missing data in the STS NCD was small and, hence, single and multiple imputation would likely give similar point estimates; (b) a slight adjustment to the standard errors would not impact the study conclusions or the published risk algorithms; (c) the large sample size would make multiple imputation less practical to implement because of long computational times.

Prior to selecting an imputation strategy, exploratory analyses were performed using CABG data from 2002 to 2003 to compare single versus multiple imputation results for predicting mortality. These analyses confirmed that the choice between single versus multiple imputation would have only a slight impact on regression coefficients. For example, the estimated odds ratio for a 5-unit increase in ejection fraction was 0.90 (with a 95% confidence interval extending from 0.83 to 0.97) under single imputation and was 0.92 (with a confidence interval extending from 0.85 to 0.99) under multiple imputation. Other variables were missing less frequently than ejection fraction and were even less sensitive to the choice between single versus multiple imputation. Additional analyses of missing data consisted of reestimating the final model coefficients using single versus multiple imputation and comparing results. A summary of these investigations, as well as model coefficients and covariance matrices, are available at www.sts.org/riskmodels. For most patients, if risk were calculated using the multiple imputation model instead of single imputation, the relative change in their risk estimate would only be 1% to 2% (eg, 5% to 5.1% is a 2% change).

Based on the considerations described above, single imputation was used with the following specific rules: (1) binary (yes/no) risk factors were modeled as yes versus no or missing. Missing data for such variables usually implies their absence, and for most binary variables the composite event rates were similar for “no” and “missing” categories; (2) missing data on categorical predictor variables were imputed to the lowest risk value, which, in most instances, was the mode. In most instances, composite event rates for patients with missing data were among the lowest. It is the policy of the STS Data Warehouse and Analysis Center to discourage missing data through this default coding practice; and (3) missing data on continuous predictor variables were imputed to the conditional median. For ejection fraction, we conditioned on congestive heart failure (CHF) and sex. For BSA, we conditioned on sex. For serum creatinine, we conditioned on renal failure (although this approach will be modified when the model is ultimately applied to version 2.61 data, as renal failure has been removed).

For model endpoints (eg, mortality), missing data were handled by modeling yes versus no or missing. Thus, cases with missing data for an endpoint were analyzed as if the endpoint did not occur. Complete case analysis was not used because “missing” was not considered to be consistently coded for these variables. For example, some STS data managers have reported that they set complications to “no” unless there is explicit documentation in the medical record that the complication occurred. Other

data managers may leave the field missing unless there is explicit documentation that the complication did not occur. Thus, missing data may reflect differences in coding practices rather than truly unknown or missing data.

Preliminary Analyses for Ordinal Categorical Variables and Continuous Variables

The QMTF conducted preliminary analyses to determine how best to model ordinal categorical variables and continuous variables. Categorical variables were entered into a logistic regression model by including a separate parameter for each category. Continuous variables were entered as piecewise linear functions (splines) with several changes of slope (knots). Terms were then removed one at a time using backward selection based on the Wald statistic. At each iteration, either two adjacent categories were collapsed into a single category or else two adjacent line segments were collapsed into a single line with no change of slope. The backward selection terminated when all adjacent categories and slopes were statistically different from one another at $p < 0.001$. This variable selection routine was performed separately for each endpoint. An expert panel determined the final coding based on the results of the backwards selection algorithm, supplemented by their clinical judgment and practical considerations. Table 3 summarizes these coding decisions.

Specific Coding Decisions

RACE AND ETHNICITY. In versions 2.35, 2.41, and 2.52.1, race was collected by choosing one of the following mutually exclusive response categories: Caucasian, black, Hispanic, Asian, Native American, and other. In version 2.61, the data collection form was modified to conform to standards adopted by the US Census Bureau. It allows for selecting one or more races per patient (ie, select all that apply), and treats ethnicity (Hispanic versus non-Hispanic) as a separate variable. Because of these differences, the mapping of race among data versions is not straightforward.

Ultimately, the QMTF decided to model race as black, Asian, Hispanic, and Caucasian/other (collapsed). Initially, these categories will be mapped to version 2.61 as follows: (1) black will include all black patients, regardless of ethnicity or additional races; (2) Hispanic will include all nonblack Hispanic patients; (3) Asian will include all Asian patients who are not also identified as black or Hispanic; and (4) all remaining patients will be placed in the Caucasian/other category. The validity of this mapping will be assessed once 2.61 data become available and future versions could employ race “bridging” methodologies.

BODY SURFACE AREA. Height and weight were replaced by BSA, which was modeled as a quadratic trend to allow for a possible U-shaped relationship with outcomes (eg, extreme obesity and cachexia). This quadratic polynomial was modeled separately for males and females. Any BSA values below 1.4 or above 2.6 were mapped to these

values respectively, which represent the approximate 1st and 99th percentiles of the empirical distribution.

ANGINA. Version 2.61 of the data collection form eliminates angina and substitutes a new variable called “cardiac presentation on admission,” within which unstable angina is one of the possible response categories. The QMTF believed that unstable angina would be coded more consistently than any other angina class, and also that this was the most important type of angina presentation to include in the models. Angina coding was therefore restricted in the new risk models to “unstable angina without MI < 7 days (yes/no).” It was necessary to exclude patients with myocardial infarction less than 7 days because the new version 2.61 does not permit simultaneous coding of angina and acute myocardial infarction.

REOPERATIVE STATUS. The most important consideration with regard to reoperative status is the number of prior sternotomies, irrespective of the specific type of procedure performed. The revised models replaced prior CABG, prior valve, and prior “other” cardiac surgery with simply the number of previous cardiovascular surgeries.

ACUITY STATUS. The new models combine resuscitation with salvage status. By definition, all salvage patients should have resuscitation coded “yes.”

NUMBER OF DISEASED CORONARY VESSELS. Outcomes are modeled using the number of diseased vessels (grouped as 0 or 1 versus 2 versus 3), as a linear effect across the three categories. This approach is consistent with the previous STS CABG models and was supported by the data.

NYHA CLASS. Version 2.61 uses NYHA class as a subfield of CHF. The grouping of NYHA IV versus less than IV (I–III) classes is consistent with all existing STS models. The final categories were no CHF, CHF not NYHA IV, and CHF plus NYHA IV.

AGE. Age was modeled as a linear spline with knots at ages 50 and 60 years.

EJECTION FRACTION. Ejection fraction (EF) was modeled linearly, and EFs below 10% and above 50% were mapped to these values respectively. Only 0.03% of patients have EFs lower than 10%; such values are considered invalid and are treated like missing data. The coding decision regarding EF values above 50% was based on preliminary analyses in which the data were used to suggest the functional form of continuous variables.

CREATININE. Creatinine was modeled as a linear spline with knots at 1.0 and 1.5. Creatinine levels less than 0.5 or greater than 5.0 were mapped to these values respectively, which represent the approximate 1st and 99th percentiles of the empirical distribution.

MORTALITY AND LENGTH OF STAY. The QMTF changed the previous STS definition of the “short postoperative length of stay (SLOS)” endpoint. The original definition did not specifically exclude early postoperative deaths, and such patients could have been inappropriately included with the remaining SLOS patients who had a particularly short and uncomplicated postoperative course. In the new models, patients who die within 5 days of surgery are included in the analysis but are not counted as a short stay.

Table 3. Final List of Candidate Variables and Coding For STS Risk Models

Candidate Variables	Coding
Continuous variables	
Age ^a	Linear spline with knots at 50 and 60.
Ejection fraction ^a	Linear; values > 50 are mapped to 50. Only 0.03% of patients have ejection fraction < 10, and that is presumed to be a data entry error; these values are considered invalid and are treated like missing data. The decision to consolidate values > 50 was based on initial exploratory analyses in which data were used to suggest the functional form of continuous variables.
Body surface area ^a	Quadratic polynomial modeled separately for males and females. Note: body surface areas < 1.4 and > 2.6 were mapped to these values, respectively. ^c
Creatinine ^a	Linear spline with knots at 1.0 and 1.5. (Only for patients not on dialysis.) Note: Creatinine values < 0.5 and > 5.0 were mapped to these values, respectively. ^d
Time trend ^a	Ordinal categorical variable with separate category for each 6-month harvest interval.
Binary variables	
Dialysis ^a	Yes/no
Preoperative atrial fibrillation ^b	Yes/no
Shock	Yes/no
Female ^a	Yes/no
Hypertension	Yes/no
Immunosuppressive treatment	Yes/no
Percutaneous coronary intervention ≤ 6 hours	Yes/no
Preoperative intra-aortic balloon pump or inotropes	Yes/no
Peripheral vascular disease	Yes/no
Unstable angina (no myocardial infarction < 7 days)	Yes/no
Left main disease	Yes/no
Aortic stenosis	Yes/no
Aortic insufficiency	Defined as at least moderate (yes/no)
Mitral insufficiency	Defined as at least moderate (yes/no)
Tricuspid insufficiency	Defined as at least moderate (yes/no)
Categorical variables	
Chronic lung disease	4 groups: (1) none, (2) mild, (3) moderate, (4) severe
CVD/CVA	3 groups: (1) no CVD, (2) CVD no CVA, (3) CVD + CVA
Diabetes mellitus	3 groups: (1) insulin diabetes, (2) noninsulin diabetes, (3) other or no diabetes
Number diseased coronary vessels	3 groups: (1) fewer than 2 diseased vessels, (2) 2 disease vessels, (3) 3 diseased vessels; modeled as linear across the categories.
Myocardial infarction	4 groups: (1) ≤ 6 hours, (2) > 6 and < 24 hours, (3) 1 to 21 days, (4) > 21 days or no myocardial infarction.
Race	4 groups: (1) black, (2) Asian, (3) Hispanic, (4) other, including Caucasian
Status	4 groups: (1) elective, (2) urgent, (3) emergent, no resuscitation, (4) salvage or emergent with resuscitation
Previous cardiovascular operations	3 groups: 0 previous, 1 previous, 2 or more previous
CHF and NYHA class	3 groups: no CHF, CHF not NYHA IV, CHF + NYHA IV
Interactions	
Age by reoperation ^a	
Age by emergent status ^a	

^a These variables were forced into each model. ^b Preoperative atrial fibrillation was forced into the model for stroke. ^c These are the approximate 1st and 99th percentiles of the empirical distribution. Values less than 1.4 were mapped to 1.4. Values greater than 2.6 were mapped to 2.6. Estimates in the extreme tails of the body surface area distribution are highly influenced by data from other regions of the body surface area distribution (owing to use of a parametric, quadratic model) and may not be reliable. ^d These are approximately the 1st and 99th percentiles of the empirical distribution. Although we used a flexible spline model, linear splines can have unreliable extreme results in the tails due to the assumption that the effect is linear above the largest knot and below the smallest knot.

CHF = congestive heart failure; CVA = cerebrovascular accident; CVD = cerebrovascular disease; NYHA = New York Heart Association.

Table 4. Discrimination of Models (C-Index)

New STS models—development sample (C-index)								
Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
0.810	0.716	0.795	0.756	0.706	0.657	0.724	0.769	0.727
New STS models—validation sample (C-index)								
Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
0.812	0.720	0.793	0.754	0.689	0.653	0.725	0.767	0.726
Old STS models—validation sample (C-index)								
Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
0.807	0.713	0.750	0.742	0.672	0.645	0.711	0.754	0.713

Comp = composite adverse outcome (any); CVA = stroke; DSWI = deep sternal wound infection; Mort = mortality; PLOS = prolonged length of stay; Reop = reoperation; RF = renal failure; SLOS = short length of stay; STS = The Society of Thoracic Surgeons; Vent = prolonged ventilation.

Final Variable Selection Procedure

Backward Selection

Using the remaining candidate variables and the coding schemes described previously, a supervised backward selection approach was then performed. Initial variable selection used the Wald χ^2 statistic with a significance criterion of 0.001. This high level of significance was chosen because of the very large sample size that resulted in quite small *p* values. An expert panel of cardiothoracic surgeons and biostatisticians then reviewed the selected variables and made several modifications. Measures of model performance (discrimination and calibration) were similar when all variables were retained in the models regardless of statistical significance or expert panel review.

Forced Variables

Several variables were included in the models regardless of statistical significance. These included all of the continuous variables (age, BSA, date of surgery [in 6-month intervals], creatinine, ejection fraction), plus sex and dialysis. In addition, atrial fibrillation was included a priori in the model for permanent stroke.

The rationale for including surgery date, a nonmodifiable variable of no intrinsic interest, was to adjust for changes in the frequency of adverse outcomes over the 5-year study period. We adjusted for surgery date to reduce potential confounding by time trends when estimating regression coefficients for variables that are of primary interest, such as preoperative clinical characteristics. For example, temporal changes in the frequency of coding for dyslipidemia, if they occur coincidentally with a secular declining trend in mortality rates, may lead to the unwarranted causal inferences unless there is adjustment for surgery date.

Date of surgery was categorized by 6-month intervals (corresponding to STS data harvests) and modeled as a linear trend across the ordinal categories. Surgery date is not included in the final risk algorithm and a patient's predicted risk is not dependent upon it. The intercept

parameter published in the Appendix has been adjusted to incorporate the time trend, and it reflects the baseline risk for a reference period of July to December 2006.

Interaction Terms

These models focused on main effects, and the final models included only four sets of preselected variable interactions: (1) sex by BSA; (2) sex by BSA squared; (3) age by reoperation; (4) age by emergent status. More extensive investigation for interactions was considered, including nonlinear, machine-learning approaches. However, the incremental value of such approaches remains uncertain [28], and interpretability can also become more problematic with numerous interaction terms.

Although multiple terms were allotted for modeling the main effects of age and reoperation, only a single degree of freedom was allotted for their interaction. The models defined a single variable interaction term for age and reoperation. It was equal to the patient's age minus 50 if the patient was at least 50 years old and had a previous CV surgery; otherwise it was equal to zero. This term represents the difference in the change of the slope of age at age 50 for patients who have had at least one previous CV surgery compared with patients who have not had a previous CV surgery. Similarly, only one degree of freedom was allotted for the interaction between age and status. The interaction represents the difference in the change of the slope of age at age 50 for patients with emergent or salvage status compared with patients with elective or urgent status. Although these interaction terms complicate the interpretation of other model variables, this was considered to be acceptable because the main focus of the analysis was prediction, not effect estimation.

Results

Model Performance

Table 4 presents the discrimination of each of the isolated CABG models as well as a comparison with the previous

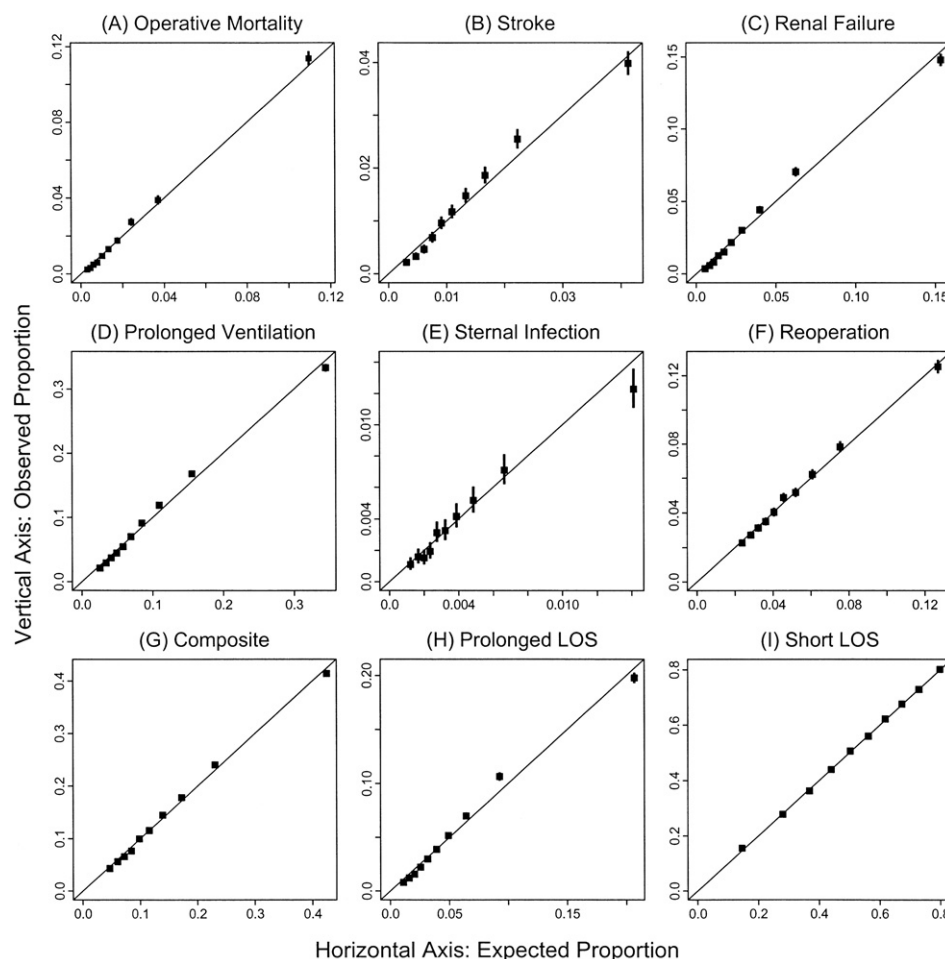


Fig 1. Plots of observed (O) versus expected (E) in validation sample

STS CABG risk models. For the new CABG models, discrimination ranged from 0.657 to 0.810 in the development sample and from 0.653 to 0.812 in the validation sample. The close agreement between c-indices from the development and validation samples reflects the large sample size and suggests that the models did not overfit the data. When the discrimination of the new and previous STS models were compared using the validation sample, the c-index of the new model was larger for each endpoint.

The Hosmer-Lemeshow test is not reported as an overall measure of calibration for these models because of its sensitivity to sample size. With samples as large as those used to develop these models, the null hypothesis will inevitably be proven false, given that all such models are only approximations [29]. As an alternative to such global measures of calibration, Figure 1 shows plots of observed versus expected event proportions within deciles of predicted risk for a variety of endpoints. For each endpoint, the absolute difference between the observed and expected proportions was less than 1.5% in each decile category. Additional analyses of model fit and discrimination are available online at www.sts.org/riskmodels.

Final Models

After calculating these measures of model performance, the final regression coefficients were estimated from the combined training and validation samples. Odds ratios for each predictor variable and model endpoint are summarized in Table 5. “Not applicable” indicates that the specific predictor was not included in a particular risk model. These final models were estimated using generalized estimating equations with empirical (sandwich) standard error estimates to account for clustering of patients within institutions [30]. An independence working correlation matrix was used to apply the generalized estimating equations method. With this approach, the estimated regression coefficients were identical to those obtained using ordinary logistic regression, but the standard errors were adjusted to account for correlated observations within hospitals.

Final Model Intercept and Coefficients

The Appendix contains the algorithm, intercept and coefficients for the final STS 2008 CABG risk models. The variance/covariance matrix is available on the web at www.sts.org/riskmodels. An on-line risk calculator is available at <http://209.220.160.181/STSTWebRiskCalc261/>.

Table 5. Estimated Odds Ratios for CABG Mortality, Morbidity, and Length of Stay Models

Variable	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
Age 60 versus 50 (no reoperation, elective)	1.36 (1.24, 1.49)	1.78 (1.58, 1.99)	1.24 (1.16, 1.33)	1.06 (1.02, 1.10)	1.43 (1.23, 1.67)	1.14 (1.09, 1.19)	1.08 (1.05, 1.11)	1.35 (1.27, 1.43)	0.77 (0.75, 0.78)
Age 70 versus 50 (no reoperation, elective)	2.53 (2.31, 2.76)	2.43 (2.19, 2.71)	1.93 (1.81, 2.07)	1.42 (1.37, 1.47)	1.70 (1.47, 1.97)	1.45 (1.39, 1.51)	1.49 (1.44, 1.53)	2.17 (2.05, 2.29)	0.44 (0.43, 0.45)
Age 80 versus 50 (no reoperation, elective)	4.70 (4.29, 5.15)	3.34 (2.99, 3.72)	3.01 (2.80, 3.24)	1.90 (1.82, 1.99)	2.02 (1.73, 2.36)	1.85 (1.76, 1.94)	2.05 (1.98, 2.12)	3.48 (3.28, 3.69)	0.25 (0.24, 0.26)
BSA 1.6 versus 2.0 among females	1.26 (1.19, 1.32)	1.15 (1.08, 1.23)	0.84 (0.80, 0.89)	1.03 (1.00, 1.06)	0.49 (0.43, 0.57)	1.23 (1.18, 1.28)	1.03 (1.01, 1.06)	0.94 (0.90, 0.97)	1.17 (1.14, 1.20)
BSA 1.6 versus 2.0 among males	1.75 (1.64, 1.86)	1.19 (1.08, 1.31)	1.24 (1.17, 1.32)	1.40 (1.34, 1.46)	0.77 (0.63, 0.93)	1.40 (1.33, 1.46)	1.35 (1.30, 1.40)	1.43 (1.36, 1.50)	0.79 (0.77, 0.82)
BSA 1.8 versus 2.0 among females	1.02 (0.99, 1.05)	1.07 (1.03, 1.11)	0.86 (0.84, 0.88)	0.95 (0.94, 0.97)	0.67 (0.63, 0.71)	1.06 (1.04, 1.08)	0.96 (0.94, 0.97)	0.90 (0.88, 0.92)	1.14 (1.13, 1.16)
BSA 1.8 versus 2.0 among males	1.20 (1.17, 1.23)	1.09 (1.05, 1.13)	1.02 (1.00, 1.04)	1.09 (1.07, 1.10)	0.85 (0.79, 0.91)	1.13 (1.11, 1.15)	1.08 (1.07, 1.09)	1.10 (1.08, 1.12)	0.96 (0.95, 0.97)
BSA 2.2 versus 2.0 among females	1.20 (1.14, 1.27)	0.95 (0.88, 1.02)	1.32 (1.27, 1.37)	1.18 (1.15, 1.22)	1.62 (1.51, 1.74)	1.03 (1.00, 1.07)	1.19 (1.16, 1.21)	1.28 (1.24, 1.32)	0.78 (0.77, 0.80)
BSA 2.2 versus 2.0 among males	1.01 (0.99, 1.03)	0.92 (0.90, 0.95)	1.17 (1.15, 1.19)	1.10 (1.08, 1.11)	1.27 (1.22, 1.32)	0.97 (0.96, 0.99)	1.07 (1.06, 1.08)	1.08 (1.07, 1.10)	0.90 (0.89, 0.91)
Creatinine 1.5 versus 1.0	1.66 (1.57, 1.76)	1.39 (1.30, 1.49)	3.36 (3.16, 3.58)	1.56 (1.51, 1.62)	1.44 (1.28, 1.62)	1.33 (1.28, 1.38)	1.76 (1.70, 1.82)	1.65 (1.59, 1.72)	0.69 (0.67, 0.71)
Creatinine 2.0 versus 1.0	1.94 (1.84, 2.04)	1.49 (1.39, 1.58)	4.06 (3.83, 4.31)	1.73 (1.68, 1.79)	1.47 (1.30, 1.65)	1.44 (1.40, 1.49)	2.05 (1.98, 2.11)	1.92 (1.86, 2.00)	0.55 (0.53, 0.57)
Creatinine 2.5 versus 1.0	2.26 (2.14, 2.39)	1.59 (1.47, 1.71)	4.90 (4.61, 5.21)	1.92 (1.85, 1.99)	1.50 (1.30, 1.72)	1.57 (1.51, 1.64)	2.39 (2.30, 2.48)	2.24 (2.15, 2.34)	0.44 (0.42, 0.46)
Dialysis versus no dialysis and creatinine = 1.0	3.84 (3.54, 4.16)	1.67 (1.48, 1.88)	NA	2.85 (2.68, 3.03)	2.13 (1.78, 2.56)	1.86 (1.73, 2.00)	2.46 (2.33, 2.60)	2.80 (2.63, 2.98)	0.27 (0.25, 0.29)
EF per 10-unit decrease	1.19 (1.17, 1.22)	1.14 (1.11, 1.16)	1.08 (1.06, 1.10)	1.18 (1.16, 1.20)	1.11 (1.07, 1.16)	1.11 (1.09, 1.13)	1.16 (1.15, 1.18)	1.17 (1.15, 1.19)	0.84 (0.83, 0.85)
Preoperative atrial fibrillation	1.36 (1.28, 1.44)	1.21 (1.12, 1.30)	1.24 (1.18, 1.30)	1.20 (1.16, 1.24)	NA	1.26 (1.21, 1.31)	1.24 (1.21, 1.28)	1.42 (1.37, 1.48)	0.61 (0.59, 0.63)
CHF not NYHA IV	1.21 (1.15, 1.28)	NA	1.36 (1.30, 1.43)	1.31 (1.26, 1.35)	1.33 (1.19, 1.48)	1.16 (1.11, 1.21)	1.27 (1.23, 1.31)	1.43 (1.38, 1.48)	0.72 (0.70, 0.75)
CHF NYHA IV	1.39 (1.31, 1.47)	NA	1.35 (1.28, 1.42)	1.52 (1.45, 1.59)	1.45 (1.25, 1.67)	1.26 (1.20, 1.32)	1.48 (1.42, 1.54)	1.50 (1.44, 1.57)	0.65 (0.61, 0.68)
Chronic lung disease, mild	1.22 (1.16, 1.29)	NA	1.14 (1.08, 1.21)	1.36 (1.31, 1.41)	1.56 (1.40, 1.73)	1.11 (1.07, 1.15)	1.23 (1.19, 1.27)	1.34 (1.29, 1.39)	0.79 (0.76, 0.82)
Chronic lung disease, moderate	1.40 (1.32, 1.49)	NA	1.25 (1.18, 1.33)	1.65 (1.57, 1.73)	1.80 (1.58, 2.06)	1.20 (1.14, 1.26)	1.42 (1.36, 1.47)	1.65 (1.58, 1.73)	0.68 (0.65, 0.71)
Chronic lung disease, severe	2.35 (2.19, 2.52)	NA	1.66 (1.54, 1.79)	2.37 (2.24, 2.51)	2.40 (2.06, 2.79)	1.54 (1.44, 1.64)	1.98 (1.90, 2.07)	2.46 (2.34, 2.60)	0.48 (0.45, 0.51)
CVD with CVA	1.31 (1.24, 1.38)	2.09 (1.96, 2.22)	1.18 (1.12, 1.23)	1.35 (1.31, 1.39)	NA	1.21 (1.17, 1.26)	1.32 (1.29, 1.36)	1.45 (1.40, 1.51)	0.70 (0.68, 0.72)

Table 5. Continued

Variable	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
CVD without CVA	1.14 (1.08, 1.20)	1.65 (1.54, 1.75)	1.11 (1.06, 1.17)	1.15 (1.11, 1.18)	NA	1.12 (1.08, 1.17)	1.17 (1.14, 1.20)	1.14 (1.10, 1.18)	0.85 (0.81, 0.89)
Diabetes, insulin dependent	1.30 (1.24, 1.37)	1.19 (1.12, 1.27)	1.80 (1.72, 1.87)	1.22 (1.18, 1.26)	2.24 (2.02, 2.48)	1.14 (1.10, 1.18)	1.30 (1.27, 1.34)	1.59 (1.53, 1.64)	0.64 (0.62, 0.66)
Diabetes, noninsulin dependent	1.01 (0.97, 1.06)	1.16 (1.11, 1.22)	1.32 (1.28, 1.36)	1.04 (1.02, 1.07)	1.38 (1.27, 1.49)	0.98 (0.96, 1.01)	1.08 (1.06, 1.10)	1.15 (1.12, 1.17)	0.87 (0.86, 0.88)
Diseased vessels (2 versus 1, or 3 versus 2)	1.17 (1.12, 1.23)	1.35 (1.29, 1.42)	1.23 (1.19, 1.27)	1.19 (1.16, 1.22)	1.15 (1.07, 1.24)	1.07 (1.05, 1.10)	1.16 (1.14, 1.18)	1.15 (1.11, 1.18)	0.81 (0.80, 0.82)
Preoperative IABP/inotropes	1.41 (1.33, 1.49)	NA	1.43 (1.36, 1.51)	2.56 (2.42, 2.72)	NA	1.37 (1.31, 1.43)	1.96 (1.86, 2.06)	1.60 (1.53, 1.67)	0.60 (0.57, 0.63)
Shock	2.29 (2.12, 2.47)	1.38 (1.23, 1.55)	1.65 (1.54, 1.77)	2.08 (1.96, 2.21)	NA	1.43 (1.34, 1.52)	2.10 (1.99, 2.23)	1.73 (1.62, 1.84)	0.58 (0.54, 0.62)
Female versus male (at BSA = 1.8)	1.31 (1.25, 1.36)	1.32 (1.24, 1.39)	1.25 (1.21, 1.31)	1.33 (1.29, 1.36)	1.19 (1.06, 1.35)	0.90 (0.87, 0.93)	1.18 (1.15, 1.21)	1.24 (1.20, 1.28)	0.65 (0.63, 0.66)
Hypertension	NA	1.29 (1.22, 1.37)	1.25 (1.20, 1.30)	1.10 (1.08, 1.13)	NA	NA	1.12 (1.10, 1.15)	1.07 (1.04, 1.11)	0.92 (0.90, 0.94)
Immunosuppressive treatment	1.48 (1.37, 1.60)	NA	1.21 (1.12, 1.31)	1.11 (1.05, 1.18)	NA	1.32 (1.24, 1.41)	1.20 (1.14, 1.26)	1.28 (1.20, 1.37)	0.80 (0.76, 0.84)
Aortic insufficiency, moderate/severe	NA	NA	NA	NA	NA	NA	NA	NA	0.82 (0.75, 0.89)
Mitral insufficiency, moderate/severe	1.31 (1.21, 1.41)	NA	NA	1.12 (1.06, 1.18)	NA	1.24 (1.16, 1.32)	1.20 (1.15, 1.26)	1.15 (1.09, 1.22)	0.85 (0.80, 0.91)
Tricuspid insufficiency, moderate/severe	NA	NA	1.31 (1.17, 1.45)	1.28 (1.18, 1.39)	NA	NA	1.24 (1.16, 1.33)	NA	0.78 (0.71, 0.87)
PCI \leq 6 hours	1.37 (1.24, 1.50)	NA	1.29 (1.16, 1.43)	1.21 (1.13, 1.29)	NA	1.30 (1.19, 1.42)	1.31 (1.23, 1.39)	1.17 (1.07, 1.27)	0.79 (0.74, 0.84)
Peripheral vascular disease	1.42 (1.36, 1.48)	1.32 (1.26, 1.39)	1.21 (1.17, 1.26)	1.22 (1.19, 1.26)	1.36 (1.24, 1.48)	1.24 (1.20, 1.28)	1.25 (1.22, 1.28)	1.31 (1.28, 1.35)	0.82 (0.81, 0.84)
Aortic stenosis	NA	NA	NA	1.18 (1.11, 1.26)	NA	NA	1.16 (1.10, 1.22)	1.15 (1.07, 1.23)	0.87 (0.82, 0.92)
Left main disease	NA	NA	NA	1.07 (1.04, 1.09)	NA	NA	1.04 (1.02, 1.06)	NA	NA
MI 1–21 days	1.37 (1.32, 1.44)	1.31 (1.25, 1.37)	1.27 (1.22, 1.32)	1.34 (1.29, 1.38)	NA	NA	1.23 (1.20, 1.25)	1.22 (1.18, 1.25)	0.88 (0.86, 0.90)
MI $>$ 6 and $<$ 24 hours	1.59 (1.46, 1.74)	1.59 (1.43, 1.76)	1.48 (1.36, 1.60)	1.59 (1.49, 1.68)	NA	NA	1.43 (1.37, 1.50)	1.31 (1.24, 1.39)	0.80 (0.76, 0.84)
MI \leq 6 hours	1.70 (1.53, 1.89)	1.49 (1.31, 1.68)	1.43 (1.29, 1.57)	1.56 (1.45, 1.67)	NA	NA	1.44 (1.35, 1.53)	1.30 (1.21, 1.40)	0.82 (0.77, 0.87)
Time trend, per 6-month harvest interval	0.97 (0.97, 0.98)	0.97 (0.96, 0.98)	1.01 (1.00, 1.02)	1.01 (1.01, 1.02)	0.97 (0.95, 0.99)	0.99 (0.99, 1.00)	1.00 (1.00, 1.01)	1.00 (0.99, 1.01)	0.99 (0.98, 1.00)
Race Asian	NA	1.33 (1.14, 1.55)	1.08 (0.96, 1.22)	1.33 (1.21, 1.47)	1.00 (0.66, 1.51)	1.31 (1.17, 1.46)	1.23 (1.15, 1.31)	1.26 (1.13, 1.40)	0.70 (0.61, 0.81)
Race black	NA	1.41 (1.30, 1.54)	1.24 (1.16, 1.33)	1.37 (1.27, 1.48)	1.30 (1.13, 1.51)	1.21 (1.14, 1.30)	1.31 (1.24, 1.38)	1.43 (1.34, 1.51)	0.69 (0.65, 0.73)
Race Hispanic	NA	1.12 (0.98, 1.27)	1.24 (1.11, 1.39)	1.16 (1.07, 1.26)	1.30 (1.07, 1.58)	1.05 (0.97, 1.13)	1.12 (1.05, 1.19)	1.09 (0.99, 1.20)	0.85 (0.77, 0.94)

Table 5. Continued

Variable	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
Reoperation, 1 previous operation ^a	3.13 (2.74, 3.57)	NA	1.52 (1.35, 1.71)	1.72 (1.58, 1.86)	NA	1.57 (1.43, 1.74)	1.61 (1.50, 1.72)	1.62 (1.47, 1.80)	0.72 (0.67, 0.78)
Reoperation, ≥ 2 previous operations ^a	4.19 (3.45, 5.09)	NA	1.58 (1.33, 1.87)	1.86 (1.62, 2.14)	NA	1.71 (1.44, 2.03)	1.84 (1.65, 2.05)	1.79 (1.53, 2.08)	0.64 (0.56, 0.73)
Status urgent ^a	1.16 (1.10, 1.22)	1.11 (1.06, 1.17)	1.12 (1.05, 1.19)	1.24 (1.18, 1.31)	1.20 (1.10, 1.32)	1.18 (1.13, 1.23)	1.18 (1.14, 1.22)	1.20 (1.15, 1.25)	0.86 (0.83, 0.90)
Status emergent, no resuscitation ^a	2.83 (2.52, 3.18)	2.12 (1.82, 2.48)	1.68 (1.49, 1.89)	2.14 (1.96, 2.34)	1.87 (1.46, 2.40)	1.83 (1.68, 1.99)	1.77 (1.64, 1.91)	2.12 (1.93, 2.32)	0.62 (0.58, 0.67)
Status emergent with resuscitation or salvage ^a	8.00 (6.91, 9.26)	2.51 (1.98, 3.18)	2.16 (1.82, 2.55)	3.01 (2.68, 3.38)	2.09 (1.45, 3.01)	2.34 (2.06, 2.65)	3.65 (3.26, 4.09)	2.39 (2.10, 2.72)	0.34 (0.30, 0.38)
Unstable angina	1.12 (1.07, 1.17)	NA	1.11 (1.05, 1.17)	1.05 (1.01, 1.10)	NA	NA	NA	NA	NA

^a Variable interacts with age. Reported odds ratio represents effect of risk factor for patients aged 50 years old.

BSA = body surface area; CHF = congestive heart failure; Comp = composite adverse outcome (any); CVA = cerebrovascular accident, or stroke; CVD = cerebrovascular disease; DSWI = deep sternal wound infection; EF = ejection fraction; IABP = intra-aortic balloon pump; MI = myocardial infarction; Mort = mortality; NA = not applicable; NYHA = New York Heart Association; PCI = percutaneous coronary intervention; PLOS = prolonged length of stay; PVD = peripheral vascular disease; Reop = reoperation; RF = renal failure; SLOS = short length of stay; Vent = prolonged ventilation.

Previously, the STS risk models were completely upgraded every 3 years, with annual recalibration in the interim to assure that the benchmark O/E ratio is always 1. In the near future, annual upgrades of the models are planned.

Limitations

Regardless of sample size or degree of statistical sophistication, all risk models are imperfect representations of reality. Although the STS risk models are based upon excellent clinical data and large sample sizes, there are some risk factors that are rare in the overall population but, when present, may be important predictors of outcome for specific patients. Some such variables, such as liver disease, are not included in the risk models, and the mortality risk for patients with these risk factors may be underestimated. Addition of a number of such variables will be considered at the next major specification upgrade.

There are other variables whose specifications undergo small but important changes over time, often in response to comments from STS database participants. These refinements are discussed on regular biweekly conference calls open to database participants, and suggested changes are regularly communicated to participants through a variety of means including FAQ's. With each major specification upgrade, they are incorporated into the new software specifications.

Audit is extremely important to assure the accuracy of any data registry. For the STS database and the risk models derived from it, robust audit is particularly critical as this registry is increasingly used for public reporting of outcomes and pay for performance. Studies suggest that the accuracy of the STS database is high for most important variables [31-35], although these audits are currently restricted to a limited number of sites annually because of budgetary constraints. In these audits, one of the most problematic variables has been 30-day mortality status (as opposed to in-hospital mortality). This is often a difficult endpoint to ascertain and may require more substantial investment of time and effort by participants, particularly for patients referred from outside their own institutions. Analysis of STS data suggests that approximately 90% of 30-day deaths occur in-hospital. Thus, if some patients recorded as being alive at 30 days have actually had their status ascertained only during the index hospitalization, the impact of this misclassification on the risk models should be negligible. This hypothesis was confirmed by comparing the odds ratios of all model variables for in-hospital versus 30-day mortality. Differences between the two were quite small, and these data are available on the web at www.sts.org/riskmodels. A new risk model for in-hospital mortality has been developed and placed on the same STS website. Furthermore, an aggressive program is in place to further enhance the accuracy of 30-day follow-up. In 2009, STS instituted a requirement that participants maintain documentation of the method by which they ascertained 30-day status, and that has become part of our routine audit. Linkage of the STS database with external death registries, such as the Social Security Death Master File, will

further support this capability. Finally, plans are being developed to expand the audit of certain key variables such as 30-day mortality to a significantly greater number of sites annually.

Conclusions

Risk-adjustment models account for the effect of patient comorbidities on outcomes. STS risk models are based upon clinical data from the STS NCD, one of the oldest and largest of all specialty registries. The value of such clinical registries is particularly evident in today's health care environment, where accreditation, regulatory compliance, reimbursement, and referrals are increasingly based upon objective data. Organizations such as the AQA and the National Quality Forum that evaluate and endorse performance measures strongly advocate the use of risk-adjusted outcomes measures.

STS believes that clinical data are superior to those derived from administrative sources. Furthermore, given the substantial implications of risk-adjusted outcomes, we believe that all risk models used for profiling quality of care should be transparent to permit comprehensive peer review and to foster credibility among stakeholders.

We present a detailed exposition of the development and validation of the 2008 STS CABG risk model. This describes not only the statistical considerations but, just as importantly, the many clinical and pragmatic judgments that are always necessary in risk model development.

References

1. Kouchoukos NT, Ebert PA, Grover FL, Lindesmith GG. Report of the Ad Hoc Committee on Risk Factors for Coronary Artery Bypass Surgery. *Ann Thorac Surg* 1988;45:348–9.
2. Clark RE. It is time for a national cardiothoracic surgical data base. *Ann Thorac Surg* 1989;48:755–6.
3. Edwards FH. Evolution of the Society of Thoracic Surgeons National Cardiac Surgery Database. *J Invasive Cardiol* 1998;10:485–8.
4. Grover FL, Shroyer AL, Hammermeister K, et al. A decade's experience with quality improvement in cardiac surgery using the Veterans Affairs and Society of Thoracic Surgeons national databases. *Ann Surg* 2001;234:464–72.
5. Shahian DM, Blackstone EH, Edwards FH, et al. Cardiac surgery risk models: a position article. *Ann Thorac Surg* 2004;78:1868–77.
6. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg* 2001;72:2155–68.
7. Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Stat Sci* 2007;22:206–26.
8. Naftel DC. Do different investigators sometimes produce different multivariable equations from the same data? *J Thorac Cardiovasc Surg* 1994;107:1528–9.
9. Iezzoni LI. "Black box" medical information systems. A technology needing assessment. *JAMA* 1991;265:3006–7.
10. Shahian DM, Hutter MM, Torchiana DF, Iezzoni LI. Transparency: a mandatory requirement for risk models. *J Am Coll Surg* 2008;206:1240–2.
11. Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association scientific statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Endorsed by the American College of Cardiology Foundation. *Circulation* 2006;113:456–62.
12. Breiman L. Statistical modeling: the two cultures. *Stat Sci* 2001;16:199–231.
13. Harrell FE Jr. Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag, 2001.
14. Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression methods in biostatistics linear, logistic, survival, and repeated measures models. New York: Springer-Verlag, 2005.
15. Jones RH, Hannan EL, Hammermeister KE, et al. Identification of preoperative variables needed for risk adjustment of short-term mortality after coronary artery bypass graft surgery. The Working Group Panel on the Cooperative CABG Database Project. *J Am Coll Cardiol* 1996;28:1478–87.
16. Tu JV, Sykora K, Naylor CD. Assessing the outcomes of coronary artery bypass graft surgery: how many risk factors are enough? Steering Committee of the Cardiac Care Network of Ontario. *J Am Coll Cardiol* 1997;30:1317–23.
17. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928–35.
18. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;27:157–72.
19. Cooper WA, O'Brien SM, Thourani VH, et al. Impact of renal dysfunction on outcomes of coronary artery bypass surgery: results from the Society of Thoracic Surgeons National Adult Cardiac Database. *Circulation* 2006;113:1063–70.
20. Grunkemeier GL, Zerr KJ, Jin R. Cardiac surgery report cards: making the grade. *Ann Thorac Surg* 2001;72:1845–8.
21. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;16:9–13.
22. Hannan EL, Kilburn H Jr, O'Donnell JF, Lukacik G, Shields EP. Adult open heart surgery in New York State. An analysis of risk factors and hospital mortality rates. *JAMA* 1990;264:2768–74.
23. Grover FL, Johnson RR, Marshall G, Hammermeister KE. Factors predictive of operative mortality among coronary artery bypass subsets. *Ann Thorac Surg* 1993;56:1296–306.
24. Grover FL, Shroyer AL, Hammermeister KE. Calculating risk and outcome: the Veterans Affairs database. *Ann Thorac Surg* 1996;62(Suppl):6–11.
25. O'Connor GT, Plume SK, Olmstead EM, et al. A regional prospective study of in-hospital mortality associated with coronary artery bypass grafting. The Northern New England Cardiovascular Disease Study Group. *JAMA* 1991;266:803–9.
26. O'Connor GT, Plume SK, Olmstead EM, et al. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery. Northern New England Cardiovascular Disease Study Group. *Circulation* 1992;85:2110–8.
27. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken: Wiley-Interscience, 2002.
28. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Stat Med* 2007;26:2937–57.
29. Marcin JP, Romano PS. Size matters to a model's fit. *Crit Care Med* 2007;35:2212–3.
30. Liang KY, Zeger SL. Longitudinal data-analysis using generalized linear models. *Biometrika* 1986;73:13–22.
31. Grover FL, Shroyer AL, Edwards FH, et al. Data quality review program: the Society of Thoracic Surgeons Adult Cardiac National Database. *Ann Thorac Surg* 1996;62:1229–31.
32. Shroyer AL, Edwards FH, Grover FL. Updates to the Data Quality Review Program: the Society of Thoracic Surgeons Adult Cardiac National Database. *Ann Thorac Surg* 1998;66:1494–7.
33. Herbert MA, Prince SL, Williams JL, Magee MJ, Mack MJ. Are unaudited records from an outcomes registry database accurate? *Ann Thorac Surg* 2004;77:1960–4.
34. Welke KF, Peterson ED, Vaughan-Sarrazin MS, et al. Comparison of cardiac surgery volumes and mortality rates

between the Society of Thoracic Surgeons and Medicare databases from 1993 through 2001. *Ann Thorac Surg* 2007;84:1538–46.

35. Welke KF, Ferguson TB Jr, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. *Ann Thorac Surg* 2004;77:1137–9.

Appendix

Regression Coefficients and Variable Definitions for STS 2008 CABG Models

For each endpoint, the formula for calculating a patient's predicted risk of the endpoint has the form:

$$\text{Predicted Risk} = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

where x_1, x_2, \dots, x_n denote patient preoperative risk factors (eg, quantitative variables such as age, and comorbidities coded as 1 = present, 0 = absent), and $\beta_0, \beta_1, \dots, \beta_n$ denote regression coefficients (numerical constants). Regression coefficients for each endpoint are presented in [Appendix Table 1](#). The variables x_1, x_2, \dots, x_n are the same for each endpoint and are defined in [Appendix Table 2](#). The regression coefficient for the time trend is not presented. Instead, the intercept has been adjusted to incorporate the time trend. This adjusted intercept reflects the baseline risk for a reference period of July to December 2006.

Appendix Table 1. Regression Coefficients

Variable	Mort	CVA	RF	Vent	DSWI	Reop	Comp	PLOS	SLOS
Intercept	-6.34090	-7.18174	-7.94605	-4.15175	-6.75378	-3.84861	-3.71671	-5.35975	2.84959
Atrial fibrillation	0.30830	0.18935	0.21351	0.17871	0.00000	0.23031	0.21565	0.35322	-0.49309
Age	-0.00259	0.00996	0.00678	0.00170	-0.00665	-0.00013	0.00247	0.00914	-0.01781
Age function 1	0.03325	0.04742	0.01496	0.00393	0.04270	0.01305	0.00515	0.02085	-0.00895
Age function 2	0.03140	-0.02582	0.02249	0.02366	-0.01895	0.01133	0.02441	0.01734	-0.02904
Age by reoperation function	-0.01714	-0.00098	-0.00291	-0.00459	0.00304	-0.00720	-0.00444	-0.00809	0.00449
Age by status function	-0.01366	-0.01363	-0.00022	-0.00106	-0.00352	-0.00435	0.00270	-0.00833	-0.00266
BSA function 1	-1.39342	-0.44041	-0.53672	-0.83950	0.65513	-0.83758	-0.75006	-0.89037	0.57952
BSA function 2	2.41303	0.06122	2.19879	2.15647	0.90025	1.16543	1.81770	2.15270	-1.83776
CHF but not NYHA IV	0.19229	0.00000	0.30971	0.26853	0.28272	0.14692	0.23695	0.35623	-0.32350
CHF and NYHA IV	0.32663	0.00000	0.30013	0.41599	0.36909	0.22846	0.39005	0.40757	-0.43827
Chronic lung disease mild	0.20273	0.00000	0.13488	0.30473	0.44371	0.10432	0.20878	0.29051	-0.23600
Chronic lung disease moderate	0.33843	0.00000	0.22530	0.50235	0.59021	0.18071	0.34720	0.50246	-0.39085
Chronic lung disease severe	0.85513	0.00000	0.50645	0.86175	0.87366	0.43034	0.68538	0.90211	-0.73862
Creatinine function 1	0.19353	0.02822	1.91934	-0.02712	-0.37465	0.01583	0.13361	-0.09060	0.00773
Creatinine function 2	0.82140	0.63174	0.50685	0.92120	1.09976	0.55107	0.99190	1.09571	-0.75781
Creatinine function 3	-0.70646	-0.52856	-2.04970	-0.68907	-0.68466	-0.39956	-0.81791	-0.70069	0.30449
CVD without prior CVA	0.13177	0.49807	0.10637	0.13792	0.00000	0.11403	0.15561	0.13271	-0.16385
CVD and prior CVA	0.26877	0.73600	0.16135	0.29946	0.00000	0.19208	0.28099	0.37248	-0.35706
Diabetes noninsulin dependent	0.01375	0.14992	0.27443	0.04283	0.31888	-0.01929	0.07453	0.13541	-0.13813
Diabetes insulin dependent	0.26312	0.17483	0.58581	0.19735	0.80627	0.12930	0.26525	0.46226	-0.44725
Dialysis	1.53777	0.54158	0.00000	1.01943	0.38312	0.63691	1.03466	0.93792	-1.30294
Ejection fraction function	0.01765	0.01274	0.00754	0.01669	0.01081	0.01063	0.01496	0.01542	-0.01756
Female	0.26801	0.27414	0.22704	0.28338	0.17792	-0.10270	0.16434	0.21488	-0.43658
Female by BSA function 1	0.82285	0.08974	0.96428	0.76954	1.11546	0.31901	0.66663	1.05623	-0.96846
Female by BSA function 2	0.05606	0.06490	-0.61086	-0.62558	0.17399	-0.02390	-0.25077	-0.35160	0.46088
Hypertension	0.00000	0.25718	0.22126	0.09930	0.00000	0.00000	0.11674	0.07200	-0.08155
IABP or inotropes	0.34193	0.00000	0.36023	0.94050	0.00000	0.31326	0.67253	0.47092	-0.51444
Immunosuppressive treatment	0.39159	0.00000	0.18881	0.10686	0.00000	0.27802	0.18030	0.24833	-0.22718
Insufficiency, aortic	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-0.19889
Insufficiency, mitral	0.26631	0.00000	0.00000	0.11169	0.00000	0.21170	0.18225	0.14174	-0.15962
Insufficiency, tricuspid	0.00000	0.00000	0.26729	0.24834	0.00000	0.00000	0.21893	0.00000	-0.24548
Left main disease	0.00000	0.00000	0.00000	0.06629	0.00000	0.00000	0.03570	0.00000	0.00000
MI 1 to 21 days	0.31810	0.27134	0.23962	0.28925	0.00000	0.00000	0.20524	0.19517	-0.12752
MI > 6 and < 24 hours	0.46614	0.46063	0.38917	0.46158	0.00000	0.00000	0.35859	0.27109	-0.22557
MI ≤ 6 hours	0.53242	0.39601	0.35421	0.44230	0.00000	0.00000	0.36337	0.26311	-0.19946
No. diseased vessel function	0.16120	0.30339	0.20729	0.17622	0.13869	0.06895	0.15075	0.13589	-0.21043
PCI ≤ 6 hours	0.31149	0.00000	0.25189	0.18695	0.00000	0.26256	0.26774	0.15633	-0.23860
Peripheral vascular disease	0.34951	0.27985	0.19308	0.20240	0.30529	0.21306	0.22277	0.27380	-0.19321
Race black	0.00000	0.34423	0.21696	0.31563	0.26572	0.19456	0.26634	0.35426	-0.37515
Race Hispanic	0.00000	0.11002	0.21645	0.14802	0.26330	0.04798	0.11289	0.08968	-0.16091
Race Asian	0.00000	0.28567	0.07579	0.28561	-0.00145	0.26855	0.20484	0.23064	-0.35049
Reop, 1 previous operation	1.13997	0.00000	0.41962	0.53987	0.00000	0.45372	0.47614	0.48534	-0.32375
Reop, ≥ 2 previous operations	1.43250	0.00000	0.45592	0.62211	0.00000	0.53695	0.61014	0.57945	-0.44745
Shock	0.82667	0.32434	0.50003	0.73290	0.00000	0.35800	0.74320	0.54575	-0.54475
Status urgent	0.14608	0.10671	0.11226	0.21738	0.18496	0.16500	0.16492	0.18202	-0.14608
Status emergent	1.04010	0.75216	0.51857	0.76090	0.62665	0.60549	0.56983	0.75083	-0.47745
Status salvage	2.07934	0.91950	0.76808	1.10085	0.73651	0.84873	1.29422	0.87072	-1.08265
Stenosis aortic	0.00000	0.00000	0.00000	0.16529	0.00000	0.00000	0.14706	0.13988	-0.14173
Unstable angina	0.11217	0.00000	0.10287	0.05060	0.00000	0.00000	0.00000	0.00000	0.00000

BSA = body surface area; CHF = congestive heart failure; Comp = composite adverse event (any); CVA = cerebrovascular accident (stroke); CVD = cerebrovascular disease; DSWI = deep sternal wound infection; IABP = intra-aortic balloon pump; MI = myocardial infarction; Mort = mortality; NYHA = New York Heart Association; PCI = percutaneous coronary intervention; PLOS = prolonged length of stay; Reop = reoperation; RF = renal failure; SLOS = short length of stay; Vent = prolonged ventilation.

Appendix Table 2. Definition of Variables Appearing in STS 2008 CABG Models

Variable	Definition
Intercept	= 1 for all patients
Atrial fibrillation	= 1 if patient has history of preoperative atrial fibrillation, = 0 otherwise
Age	= Patient age in years
Age function 1	= max (age–50, 0)
Age function 2	= max (age–60, 0)
Age by reop function	= Age function 1 if surgery is a reoperation, = 0 otherwise
Age by status function	= Age function 1 if status is emergent or salvage, = 0 otherwise
BSA function 1	= max (1.4, min [2.6, BSA]) – 1.8
BSA function 2	= (BSA function 1) ²
CHF but not NYHA IV	= 1 if patient has CHF and is not NYHA class IV, = 0 otherwise
CHF and NYHA IV	= 1 if patient has CHF and is NYHA class IV, = 0 otherwise
CLD mild	= 1 if patient has mild chronic lung disease, = 0 otherwise
CLD moderate	= 1 if patient has moderate chronic lung disease, = 0 otherwise
CLD severe	= 1 if patient has severe chronic lung disease, = 0 otherwise
Creatinine function 1	= max (0.5, min [creatinine, 5.0]) if patient is not on dialysis, = 0 otherwise
Creatinine function 2	= max ([creatinine function 1] – 1.0, 0)
Creatinine function 3	= max ([creatinine function 1] – 1.5, 0)
CVD without prior CVA	= 1 if patient has history of CVD and no prior CVA, = 0 otherwise
CVD and prior CVA	= 1 if patient has history of CVD and a prior CVA, = 0 otherwise
Diabetes, noninsulin	= 1 if patient has diabetes not treated with insulin, = 0 otherwise
Diabetes, insulin	= 1 if patient has diabetes treated with insulin, = 0 otherwise
Dialysis	= 1 if patient requires dialysis preoperatively, = 0 otherwise
Ejection fraction function	= max (50 – ejection fraction, 0)
Female	= 1 if patient is female, = 0 otherwise
Female by BSA function 1	= BSA function 1 if female, = 0 otherwise
Female by BSA function 2	= BSA function 2 if female, = 0 otherwise
Hypertension	= 1 if patient has hypertension, = 0 otherwise
IABP or inotropes	= 1 if patient requires IABP or inotropes preoperatively, = 0 otherwise
Immunosuppressive treatment	= 1 if patient given immunosuppressive therapy within 30 days, = 0 otherwise
Insufficiency, aortic	= 1 if patient has at least moderate aortic insufficiency, = 0 otherwise
Insufficiency, mitral	= 1 if patient has at least moderate mitral insufficiency, = 0 otherwise
Insufficiency, tricuspid	= 1 if patient has at least moderate tricuspid insufficiency, = 0 otherwise
Left main disease	= 1 if patient has left main disease, = 0 otherwise
MI 1 to 21 days	= 1 if history of MI 1 to 21 days prior to surgery, = 0 otherwise
MI > 6 and < 24 hours	= 1 if history of MI >6 and <24 hours prior to surgery, = 0 otherwise
MI ≤ 6 hours	= 1 if history of MI ≤ 6 hours prior to surgery, = 0 otherwise
No. diseased vessel function	= 2 if triple-vessel disease, = 1 if double-vessel disease, = 0 otherwise
PCI ≤ 6 hours	= 1 if patient had PCI ≤ 6 hours prior to surgery, = 0 otherwise
Peripheral vascular disease	= 1 if patient has peripheral vascular disease, = 0 otherwise
Race black	= 1 if patient is black, = 0 otherwise
Race Hispanic	= 1 if patient is nonblack Hispanic, = 0 otherwise
Race Asian	= 1 if patient is nonblack, non-Hispanic, and is Asian, = 0 otherwise
Reop, 1 previous operation	= 1 if patient has had exactly 1 previous CV surgery, = 0 otherwise
Reop, ≥ 2 previous operations	= 1 if patient has had 2 or more previous CV surgeries, = 0 otherwise
Shock	= 1 if patient was in shock at time of procedure, = 0 otherwise
Status urgent	= 1 if status is urgent, = 0 otherwise
Status emergent	= 1 if status is emergent (but not resuscitation), = 0 otherwise
Status salvage	= 1 if status is salvage (or emergent plus resuscitation), = 0 otherwise
Stenosis aortic	= 1 if patient has aortic stenosis, = 0 otherwise
Unstable angina	= 1 if patient has unstable angina, no MI within 7 days of surgery, = 0 otherwise

BSA = body surface area; CHF = congestive heart failure; CLD = chronic lung disease; CVA = cerebrovascular accident, or stroke; CVD = cerebrovascular disease; DSWI = deep sternal wound infection; EF = ejection fraction; IABP = intra-aortic balloon pump; MI = myocardial infarction; Mort = mortality; NYHA = New York Heart Association; PCI = percutaneous coronary intervention; PLOS = prolonged length of stay; Reop = reoperation; Comp = composite adverse event (any); RF = renal failure; SLOS = short length of stay; STS = The Society of Thoracic Surgeons; Vent = prolonged ventilation.

Developer Response to Scientific Methods Panel’s Preliminary Analysis

Measure Number: 0696

Measure Title: STS CABG Composite Score

Measure Developer/Steward: Society of Thoracic Surgeons

[Remove any bullet that is not needed or does not apply to your response; add additional bullets for issues identified and responses as needed. Each issue identified should be addressed separately.]

Validity

- **Issue 1:**

SMP members expressed concerns related to STS methodology for demonstrating empirical validity and content validity. (e.g. “An association with a different construct that is expected to be correlated with measure 0696 was not assessed... Could the developers demonstrate that the scores are associated with another related measure?”)

- **Developer Response 1:**

With most individual measures, it is possible to find another external quality metric against which to assess validity. In the case of the STS CABG composite, we have purposely included all the major quality metrics that have been used for CABG. Thus, they are within the composite and are not available as separate external measures for validation. That is the reason we showed the correlation of the overall composite score with results for each of the domains.

- **Issue 2:**

An SMP member suggested that the STS risk adjustment model relevant to measure 0696 needs to be updated.

- **Developer Response 2:**

The STS updated and published the relevant risk models in 2018;* please see “STS 2018 Adult Cardiac Surgery Risk Models: Part 2” attached. (The “Part 1” paper is also provided for additional background on the risk model updates.) Please advise if you would like us to revise 2b3.4a in the Composite Measure Testing form for this measure with the updated list of risk factors.

* We did not previously include the updated risk models in our measure documentation due to 2015 NQF guidance

(<http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80308>) describing “decreased emphasis” in the endorsement maintenance process on updated reliability and validity (including risk adjustment) information: “If prior testing adequate, no need for additional testing at maintenance...” We also did

not wish to create a mismatch between the date of our risk adjustment model and the dates of various data analyses provided in our Composite Measure Testing form.

- **Issue 3:**

SMP members requested an update to the data used to demonstrate validity (2b1.3).

- **Developer Response 3:**

It is not feasible for the STS to update the analyses in 2b1.3 within the timeframe specified (by 10 AM ET on Oct. 16). Given that the SMP will not be meeting until Oct. 28-29, we will appreciate an extension to your deadline for this information, i.e. a due date closer to the SMP meeting date.

The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1—Background, Design Considerations, and Model Development



David M. Shahian, MD, Jeffrey P. Jacobs, MD, Vinay Badhwar, MD, Paul A. Kurlansky, MD, Anthony P. Furnary, MD, Joseph C. Cleveland, Jr, MD, Kevin W. Lobdell, MD, Christina Vassileva, MD, Moritz C. Wyler von Ballmoos, MD, PhD, Vinod H. Thourani, MD, J. Scott Rankin, MD, James R. Edgerton, MD, Richard S. D'Agostino, MD, Nimesh D. Desai, MD, PhD, Liqi Feng, MS, Xia He, MS, and Sean M. O'Brien, PhD

Department of Surgery and Center for Quality and Safety, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts (DMS); Division of Cardiac Surgery, Johns Hopkins University School of Medicine, Baltimore, Maryland (JPJ); Division of Cardiovascular Surgery, Johns Hopkins All Children's Heart Institute, St. Petersburg, Florida (JPJ); Department of Cardiovascular and Thoracic Surgery, West Virginia University, Morgantown, West Virginia (VB, JSR); Division of Cardiac Surgery, Columbia University, New York, New York (PAK); Starr-Wood Cardiothoracic Group, Portland, Oregon (APF); Division of Cardiothoracic Surgery, University of Colorado Anschutz School of Medicine, Aurora, Colorado (JCC); Atrium Health, Cardiovascular and Thoracic Surgery, Charlotte, North Carolina (KWL); Division of Cardiac Surgery, University of Massachusetts Medical School, Worcester, Massachusetts (CV); Houston Methodist DeBakey Heart and Vascular Center, Houston, Texas (MCWvB); Department of Cardiac Surgery, MedStar Heart and Vascular Institute, Georgetown University, Washington, DC (VHT); The Heart Hospital Baylor Plano, Plano, Texas (JRE); Division of Thoracic and Cardiovascular Surgery, Lahey Hospital and Medical Center, Burlington, Massachusetts (RSD); Division of Cardiothoracic Surgery, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania (NDD); and Duke Clinical Research Institute, Duke University Medical Center, Durham, North Carolina (LF, XH, SMO)

Background. The last published version of The Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database (ACSD) risk models were developed in 2008 based on patient data from 2002 to 2006 and have been periodically recalibrated. In response to evolving changes in patient characteristics, risk profiles, surgical practice, and outcomes, the STS has now developed a set of entirely new risk models for adult cardiac surgery.

Methods. New models were estimated for isolated coronary artery bypass grafting surgery (CABG [$n = 439,092$]), isolated aortic or mitral valve surgery ($n = 150,150$), and combined valve plus CABG procedures ($n = 81,588$). The development set was based on July 2011 to June 2014 STS ACSD data; validation was performed using July 2014 to December 2016 data. Separate models were developed for operative mortality, stroke, renal failure, prolonged ventilation, reoperation, composite major morbidity or mortality, and prolonged

or short postoperative length of stay. Because of its low occurrence rate, a combined model incorporating all operative types was developed for deep sternal wound infection/mediastinitis.

Results. Calibration was excellent except for the deep sternal wound infection/mediastinitis model, which slightly underestimated risk because of higher rates of this endpoint in the more recent validation data; this will be recalibrated in each feedback report. Discrimination (c-index) of all models was superior to that of 2008 models except for the stroke model for valve patients.

Conclusions. Completely new STS ACSD risk models have been developed based on contemporary patient data; their performance is superior to that of previous STS ACSD models.

(Ann Thorac Surg 2018;105:1411–8)

© 2018 by The Society of Thoracic Surgeons

Although recalibrated each harvest since their development in 2008, the last published version of The Society of Thoracic Surgeons (STS) Adult Cardiac

Surgery Database (ACSD) risk models was based on patient data from 2002 to 2006. To incorporate evolving changes in patient characteristics, risk profiles, surgical practice, and outcomes, the STS has developed a set of entirely new risk models for adult cardiac surgery.

In this two-part report, we present the 2018 STS adult cardiac surgery risk models. Part 1 provides an introductory background regarding the history of STS risk modeling, general principles used to develop these

Accepted for publication March 9, 2018.

The STS Executive Committee approved this document.

Address correspondence to Dr Shahian, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114; email: dshahian@partners.org.

Abbreviations and Acronyms

ACSD	= Adult Cardiac Surgery Database
AVR	= aortic valve replacement
CABG	= coronary artery bypass grafting surgery
DSWI	= deep sternal wound infection/mediastinitis
O/E	= observed to expected ratio
PLOS	= postoperative length of stay
SES	= socioeconomic status
SLS	= significance level to stay
STS	= The Society of Thoracic Surgeons

models, and a systematic explanation of the model development process. Part 2 [1] provides more extensive technical detail regarding the statistical methodology and results.

Genesis of STS Risk Models—Federal Government “Death Lists”

The modern era of transparency and accountability in health care began not with the Affordable Care Act but more than 2 decades earlier, in March 1986, with the publication of essentially unadjusted hospital-level mortality data by the Health Care Financing Administration, the predecessor of the Centers for Medicare and Medicaid Services. One of the specific, high-profile procedures targeted was coronary artery bypass grafting surgery (CABG), leading hospitals and surgeons to complain that the inherent risk of their patients was not being considered in these so-called Medicare “death lists.”

STS Ad Hoc Committee on Risk Factors

In response, STS formed an Ad Hoc Committee on Risk Factors chaired by Dr Nick Kouchoukos. This committee issued a Statement of Concern in October of the same year, followed by a formal report strongly advocating for risk adjustment when profiling provider performance [2]. The authors of the report correctly noted that unadjusted rates were misleading to the public as they did not account for preoperative patient severity and acuity. They suggested that this might lead providers to avoid treating high-risk patients, anticipating the phenomenon we now refer to as risk aversion [3, 4]. They advised that risk models and performance measures should focus on relatively homogeneous procedure categories, such as isolated CABG. Combined procedures (eg, CABG plus aortic valve replacement [AVR]) should be considered in separate models as they have higher inherent risk. Finally, they presciently noted that risk-adjusted mortality is an important but inadequate metric by which to assess performance; comprehensive performance measurement should also include risk-adjusted rates of other complications such as reoperation or stroke. This concept was ultimately realized in the development of STS composite performance measures [5–11]. Thirty years later,

these principles remain fundamental tenets of all STS risk models and performance measures.

Risk Models Require Optimal Data—Origins of STS National Database

A prerequisite for the development of early STS risk adjustment models was the availability of clinically granular, standardized data to characterize preoperative patient comorbidities. Because such data were unavailable except for idiosyncratic institutional registries or research datasets for one-time studies, it was recognized that a national, clinical data registry for cardiothoracic surgery was needed. This was the proximate stimulus for development of the STS National Database in 1989 [12].

Early STS Risk Models

During the same period in the late 1980s, Dr Fred Edwards had begun to explore health care applications of Bayesian techniques, initially for diagnostic evaluation and soon thereafter for risk prediction in cardiac surgery. Initial studies were limited to single institution data, but in 1994, Edwards, Clark, and Schwartz published their landmark article on the application of these Bayesian approaches to cardiac surgery risk adjustment, using data from the nascent STS National Database [13]. Subsequently, logistic risk models were introduced and have been used in subsequent risk model iterations [13–21].

Expanding Family of STS Risk Models

This portfolio of risk models has expanded from its original focus on one procedure (CABG) and one outcome (mortality) to include other major cardiothoracic procedures—valve replacement, congenital heart surgery, and general thoracic surgery—as well as other endpoints such as complications and length of stay. These models have been used to provide risk adjustment for a growing family of STS performance metrics, including composite measures [5–11], many of which are publicly reported [22, 23].

Each adult cardiac surgery risk model is published in the peer-reviewed literature, and relevant intercepts and coefficients are publicly available either in peer-reviewed journals or on the STS or Duke Clinical Research Institute websites. As part of each data harvest, each model is recalibrated so that the expected number of events exactly matches the observed number during that harvest, resulting in an observed to expected (O/E) ratio of 1. Periodically these models are completely revised, as described in this report, to include new risk factors and to reflect changes in cardiothoracic practice and outcomes.

Risk Model Applications

Risk models have many potential applications. For example, they define a standard default set of covariates for STS research analyses, they provide risk scores for real-time patient counseling and shared decision making, and they can inform and document performance improvement initiatives. However, the primary motivation in developing the current risk models was to provide

robust case-mix adjustment for estimating risk-adjusted outcomes, which are subsequently used in STS feedback reports and voluntary public reporting. Although parsimony and ease of use were secondary considerations, predictive accuracy was the paramount goal that guided model development.

Appropriate Interpretation of Risk-Adjusted Outcomes

The proper interpretation of risk-adjusted outcomes rates or O/E ratios is crucial [24]. Because of technical considerations, including the large number of clinical covariates compared with relatively few strata of interest in most epidemiologic studies, health care risk models such as those used by the STS almost always use indirect rather than direct standardization. In direct standardization, rates from various strata (eg, age) within the study population of interest (eg, in the case of profiling, a specific hospital) are applied to a reference or standard population. That may allow direct comparisons of the standardized rates from different study populations, as they have all been applied to the same standard population.

In indirect standardization, the reverse process is carried out. Rates derived from the benchmark population for characteristics of interest are applied to the study population (in the case of profiling, a health care provider's patients) to obtain their so-called expected outcomes or rates. Each hospital's indirectly standardized rate of an endpoint or their O/E ratio are based only on the patients for whom it cares, and its case mix may be quite different than that of another specific hospital. There may be patients at one hospital for whom there is no analogue at another hospital.

Because of these differences, it would be inappropriate to compare indirectly standardized rates or ratios between specific providers. For example, consider one provider caring mainly for low-risk patients and another caring predominately for high-risk patients. The indirectly standardized outcomes at the former may not include any of the high risk-patients seen at the latter. An O/E ratio of 0.8 achieved by a hospital caring for low-risk patients gives no assurance that this hospital could achieve similar results if confronted with higher risk patients. These concepts have critical implications for proper interpretation, yet they are often misunderstood or ignored [24].

It is most appropriate to interpret indirectly standardized outcomes or O/E ratios, including those from STS risk models, as representing a provider's actual results for their specific patients compared with what would have been expected (technically referred to as the counterfactual outcomes) for the same patients based on the performance of all providers who contributed to the benchmark population. The STS and others often classify these results as worse than expected, as expected, or better than expected performance.

For the same reason, rating of providers (eg, better or worse than expected or as expected) compared with a national benchmark is appropriate, whereas rankings (no. 1, no. 2, and so forth, which implies that hospital 1 is better than hospital 2) are not.

Considerations in Risk Model Development

This 2018 complete update of STS adult cardiac surgery risk models by the STS Quality Measurement Task Force, spanning several years of work by surgeons and statisticians, was based on a number of important considerations.

Why Risk Adjust?

Although the need for risk models may seem axiomatic, it is worth remembering that some health care performance measures still lack adequate risk adjustment. Outcomes measures should be risk adjusted whenever there are important patient characteristics that significantly affect the outcome of interest, and when the prevalence of such factors varies across providers. Both these conditions are satisfied in cardiac surgery [25]. Robust risk adjustment also enhances provider acceptance of outcomes measures and helps to mitigate risk aversion.

Target Procedures

Selection of a target procedure or population for risk-adjusted outcomes involves a tradeoff between adequate sample size (which might argue for broader procedure categories) and clinically coherent, more homogeneous cohorts, which have greater face validity and specificity [25]. Because of the large numbers of all procedures available in the STS ACSD, the Quality Measurement Task Force created separate risk models for the most commonly performed adult cardiac surgical procedures: isolated CABG, isolated AVR, isolated mitral valve repair or replacement, AVR plus CABG, and mitral valve repair or replacement plus CABG.

Endpoint Selection

The nine outcomes we studied—operative mortality, stroke, renal failure, prolonged ventilation, reoperation, mediastinitis/deep sternal wound infection (DSWI), major morbidity or mortality composite, prolonged post-operative length of stay (PLOS), or short PLOS—were chosen based on historical precedent, clinical impact, resource use, and inclusion in current performance metrics (eg, STS composite scores). We created separate risk models for each outcome and procedure except DSWI. Because its incidence is quite low (generally less than 0.5%), separate DSWI models for any one procedure are unreliable and subject to overfitting. Accordingly, as described in Part 2 of this report, we created a single model for DSWI that encompassed all the major procedures, with an indicator variable for the specific procedure. With the potential exception of CABG using bilateral internal thoracic arteries, the major DSWI risk factors (eg, diabetes mellitus, obesity, immunosuppression, severe chronic lung disease) were not expected to vary dramatically across procedures.

Socioeconomic Indicators

Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic status (SES) or sociodemographic

factors (eg, race, ethnicity, education, income, payer [eg, Medicare-Medicaid dual eligible status]) is a topic of intense health policy debate [26]. Some argue that in the absence of adjustment for these variables, the outcomes of hospitals that care for a disproportionate percentage of low SES patients will be unfairly disadvantaged, perhaps leading to financial or reputational penalties. Opponents argue that inclusion of SES factors in risk models may “adjust away” disparities in quality of care, and they advocate the use of stratified analyses instead. Also, readily available SES factors have often not demonstrated significant impact on outcomes, perhaps because they are not sufficiently granular or relevant. Finally, even SES proponents agree that these factors make more sense conceptually for some outcomes (eg, readmission) than for others (hospital mortality, complications). Notably, as part of an National Quality Forum pilot project, the STS specifically studied dual eligible status in the STS readmission measure [27] and found minimal impact.

In developing the new STS risk models, we avoided these more philosophical and health policy arguments regarding SES adjustment and based our modeling decisions on empiric findings and consideration of the model’s primary intended purpose—optimal case mix adjustment. Conceptually, our goal was to adjust for all preoperative factors that are independently and significantly associated with outcomes and that vary across STS participants. For example, race will continue to be in our risk models as it has been previously, but not conceptually as a SES indicator. Race has an empiric association with outcomes and has the potential to confound the interpretation of a hospital’s outcomes, although we do not know the underlying mechanism (eg, genetic factors, differential effectiveness of certain medications, rates of certain associated diseases such as diabetes and hypertension, and potentially SES for some outcomes such as readmission).

Interaction Terms

Consistent with previous STS risk models, to facilitate interpretability we elected to focus on main effects and did not include many interaction terms. In general, we allowed the effect of each patient factor to differ depending on the type of operation but not on other patient factors.

Parsimonious Versus Nonparsimonious Models

A fundamental decision in risk modeling is how many risk variables will be included. There are cogent arguments for both parsimonious and more expansive models. Most of the predictive power of risk models is contained in a relatively few important covariates [28, 29], and addition of more variables usually does not substantially change the c-index or area under the receiver-operating characteristics curve, a common (although not necessarily the best) measure of model performance. One often-used rule of thumb proposed by statistician Frank Harrell is that at least 10 endpoints are required for each variable in a model [30].

Parsimonious models are computationally simpler, faster, and more likely to converge. They are easier to run in production mode, easier for software vendors when changes or updates are implemented, and easier for practicing clinicians to use when providing patients with their estimated risk of surgery (ie, they only need to enter a small number of variables into the STS online risk calculator). An excessive number of predictors (over-parameterized model) for the number of available endpoints can lead to unstable, noisy estimates in which different random samples from the same population could produce markedly different results. They may overfit the results to the specific data used for model development but the model may not generalize well to other or subsequent populations. Finally, although our risk models are built using 3 years of data, these models will typically be utilized with smaller samples, often 1 year of data. A highly parameterized model might run adequately in the development set, but not in production mode with smaller numbers of patients.

However, there are also disadvantages to highly parsimonious models, including face validity. These models, which sometimes include only a few variables, may have reasonable overall performance at the population level. However, they necessarily exclude variables that, although uncommon, are highly predictive of adverse outcomes when present, such as severe liver disease. When these risk factors are present, the predictive accuracy of highly parsimonious models may be compromised, which would violate the guiding principle we followed in developing these models. Specifically, the risk of patients with high-impact features not adjusted for in the model will be underestimated. Failure to adjust for such high-risk characteristics could lead clinicians to avoid caring for patients with these risk factors, as their severity would not be accounted for in their risk-adjusted outcomes [3, 4].

We have opted for a middle ground: model building using backward selection from a full model, a process during which we attempted to optimize both clinical face validity and statistical performance. Using the process described below, surgeons and statisticians selected the most parsimonious models possible that did not exclude any clinically critical variables or significantly compromise predictive accuracy.

The 2018 STS Adult Cardiac Surgery Risk Models—General Principles

STS Database Version Mapping

The 2008 STS ACSD risk models were developed using data from STS versions 2.35, 2.41, and 2.52 and were designed for use with the concomitantly released version 2.61. The new 2018 risk models were developed using data from version 2.73 and tested using version 2.81. To be a viable candidate for the new 2018 models, a variable had to exist in version 2.73 and that same variable or a closely related, mappable analogue needed to be present in version 2.81. This was particularly challenging for

certain variables, such as preoperative atrial fibrillation, in which the categories or choices for specific variables (in technical terms, the parameterization) has changed between versions.

Exploratory Analyses of Candidate Variables

After defining the source data for the new risk models and performing any required mapping of variables across data versions, initial exploratory analyses of all potential candidate variables (those included in the earlier 2008 risk models plus new variables added in STS ACSD version 2.73) was performed. For each candidate variable, procedure, and outcome, we examined the percentage of patients in each group of a categorical variable; the mean and median value for continuous variables; the percentage missing data (or test not performed); the bivariate associations of the candidate variable with outcomes; and reliability estimates for full and reduced models for various outcomes.

Certain variables were retained as candidates for some scenarios but not others. For example, the use of angiotensin-converting enzyme inhibitors and angiotensin receptor-blocking agents may lead to postoperative renal dysfunction or vasoplegia, but continuing or discontinuing these agents preoperatively is based on surgeon judgment, making it a suboptimal variable for case-mix adjustment in elective cases. Conversely, surgeons may not have the option to consider discontinuing these agents for urgent or emergent cases, and they then become reasonable candidate variables for those scenarios.

Other risk factors known to effect individual patient outcomes, such as pulmonary hypertension, were, after extensive discussions, not included as candidate variables owing to temporal inconsistencies of parameter measurement (interventional laboratory versus operating room), potential acute alterations with intravenous fluid or drug therapy, and a high proportion of missing data.

Missing Data

The frequency of missing data was less than 1% for most preprocedural variables. We excluded from further consideration those few variables with missing data rates greater than 5%, or variables reflecting a test or study that had not been performed in more than 5% of the relevant study population. Examples of excess missing data precluding their use in modeling included bilirubin (missing 20%) and international normalized ratio (missing 8%), which prevented modeling of the Model for End-Stage Liver Disease score; hemoglobin A1c (missing 21%), an important marker of diabetes; etiology of valvular disease (missing in more than 10% in each valve population); and 5-m walk test (missing or not performed in 95% of patients). Previous studies have shown that the latter, when abnormal (greater than 6 seconds), increases risk twofold to threefold [31, 32]. This information regarding excluded variables is an important reminder to STS ACSD participants that complete data are essential to optimize risk model development.

Imputation strategies for the initial exploratory and variable selection analyses, and for reestimation of covariate regression coefficients in the final model, are discussed in Part 2 of this report.

Feasible Number of Candidate Predictors

Initial review identified more than 50 preoperative variables (and more than 100 potential parameters) for possible inclusion in the various models. To empirically assess the theoretical limitations posed by including such a large number of candidate variables (and their associated subcategories), we used bootstrap simulations to estimate a measure of signal-to-noise ratio reliability [33]. Here, signal refers to the amount of variation in risk across patients, whereas noise refers to the amount of error in the estimation of each patient's risk. We reasoned that the amount of acceptable noise depends in part on the magnitude of signal. For example, an average estimation error of ± 1 percentage point may be acceptable if true risk ranges from 0% to 50% across patients but unacceptable if true risk ranges from 0% to 2%.

A complete description of the methods and results of these calculations are presented in Part 2 of this report. Briefly, we found that the average estimation error was lower for CABG models than for valve or valve plus CABG models, mainly because of larger sample sizes. The most striking finding was that estimation error was consistently quite high for the sternal infection models for valve and valve plus CABG, as this is the least common endpoint (average prevalence 0.3%). That led to a decision to combine sternal infection results for all procedures into one model to increase the effective number of endpoints, as described above.

Surgeon Perspectives

Development of the new STS risk models was largely data driven and avoided forced, subjective inclusion or exclusion of variables. However, in addition to their active participation in the evaluation of the data analytics, surgeons did provide clinical insights at various stages of the development process. For example, as part of the initial exploratory analyses, surgeon members of the Quality Measurement Task Force assessed each variable's clinical importance for inclusion in the risk model, assigning it a rating of 1 to 10. Some variables (use of tobacco products other than cigarettes; dyslipidemia) were thought a priori to have little clinical relevance, had not been included in previous models, or demonstrated minimal association with outcomes in bivariate analyses. Given that the total number of variables we could include was limited, these variables were not considered further. Conversely, as described subsequently, some variables were considered so important for face validity that the *p* value for backward selection was largely based on the ability to retain these variables.

Optimal Coding Efficiency

Before final model selection, we determined the most efficient coding, or parameterization, of candidate variables, typically by collapsing or combining clinically

related or collinear variables, which often had similar magnitude associations with specific endpoints. Some variables had computationally excessive categories, or the categories had changed between versions, such as preoperative myocardial infarction, arrhythmias, and payer. In other instances, uncommon but important variables were combined with other related variables—for example, our decision to code catheter support devices and extracorporeal membrane oxygenation as shock, because that is their usual indication. Race and ethnicities have multiple categories and potential combinations—we chose to parameterize these as we had in the past, with six major categories. Similarly, although there are many payer options available for coding in the database, for modeling purposes we reduced these to a limited number of broad, coherent categories.

Other coding issues were similarly challenging. For example, body mass index and body surface area measure slightly different characteristics of body habitus and shape although both use the same height and weight inputs. After considerable discussion, both body mass index and body surface area were included as candidate variables, subject to further winnowing during the backward selection process.

For continuous variables, additional exploratory analyses helped determine how best to model the association of the variable with specific outcomes (eg, linear, quadratic, polynomial, splines).

Competing Risks

All nonfatal endpoints and complication analyses have the potential for bias due to the competing risk of death. If a patient dies on day 1 postoperatively, that patient does not have the opportunity to have other complications such as prolonged ventilation. Fortunately, the competing risk—mortality—occurs relatively uncommonly compared with most nonfatal endpoints. In most instances, risk models for complications of medical or surgical treatment have not considered the competing risk issue, other than to point out that the nonfatal endpoint (eg, readmission [34]) and mortality should both be examined or even combined in a composite, such as risk-adjusted mortality and morbidity in all STS adult cardiac surgery composite measures [5–11].

For several outcomes in the new risk models, the issue of competing risk was extensively discussed. For example, what patients should be included in the numerator and denominator of the short PLOS measure? In some instances, short PLOS may result from early death, so it could be argued those patients who die less than 6 days after surgery should be excluded from the denominator of this measure—in other words, they should be ineligible to receive credit for this measure as short PLOS is viewed as a favorable outcome. However, with a smaller denominator, the proportion of that hospital's short PLOS patients will appear larger (better), in effect “rewarding” hospitals for early mortality. Weighing the pros and cons of the various approaches, we decided that it was best to retain all patients in the denominator, irrespective of why they are discharged early, but not give numerator credit

for a short PLOS to those patients who died at less than 6 days.

The prolonged PLOS (more than 14 days) measure presents similar issues, as it effectively gives credit for patients who die less than 14 days postoperatively. For example, a patient who dies on day 8 postoperatively will not be in the numerator, as their length of stay is not more than 14 days, which would appear to be favorable for the hospital. One option would be to change the measure numerator statement to both more than 14 days and discharged alive, but that would change the meaning of a longstanding STS measure and might be misleading for patients and other stakeholders. For example, a hospital might have a very low (favorable) prolonged PLOS because patients are efficiently managed and rarely require longer stays; alternatively, prolonged PLOS might be low because more of their patients die in hospital and therefore do not meet the numerator criteria. If, however, inhospital deaths are removed from the denominator, the apparent proportion of long hospital stays will increase. Therefore, a site that has better salvage rate from serious complications (lower failure to rescue) will erroneously appear to be worse performing (ie, prolonged PLOS numerator will be relatively larger in proportion to the reduced denominator).

After weighing all these possibilities, we adopted the following definitions. Short PLOS is discharge alive within 6 days of surgery—all patients are in the denominator but there is no numerator credit for patients who die less than 6 days after surgery. The prolonged PLOS measure denominator encompasses all patients, including those who died less than 14 days after surgery; all patients who are discharged at more than 14 days are included in the numerator, including nonsurvivors. These PLOS outcomes are primarily measures of resource use and efficiency and should always be viewed in combination with clinical balancing measures (in this case, mortality) or as part of a composite measure.

Final Model Selection

After candidate covariates had been chosen for each combination of population and endpoint, we systematically selected final model variables and estimated their associated coefficients. Separate, comprehensive analyses were considered for each major procedure group (CABG, valve, valve plus CABG). Backward selection was used for model development, as it is computationally efficient and provides the ability to compare the results of full and reduced (more parsimonious) models.

When selecting the final model, determining the optimal significance level to stay (SLS) in backward selection was the primary focus, as discussed in detail in Part 2 of this report. Historically, this has typically been decided a priori, choosing a “standard” level such as $p = 0.05$. Rather than using such a predetermined, identical value for each procedure and endpoint, we elected to make these decisions based on empirical data and face validity. Surgeons and statisticians first examined the empirical data (see Part 2) for each SLS: a variety of

statistical tests of model performance; internal and external (cross validated) calibration plots for the overall population cohort and selected subgroups; and the number of variables and parameters (including categories of variables) retained at each level.

To reduce the number of model comparisons, the SLS evaluation process was limited to six possible values: $SLS = 0.0001$, $SLS = 0.001$, $SLS = 0.01$, $SLS = 0.05$, $SLS = 0.1$, and $SLS = 1$, where $SLS = 1$ means that all candidate covariates are retained (ie, the full model, no backward selection).

For each population, endpoint, and SLS value, the following information was reviewed by the modeling committee, and will be discussed in more detail in Part 2 of this article:

- Optimism-adjusted c-statistic
- Optimism-adjusted slope
- Calibration plots of observed versus expected outcome by decile of predicted risk, overall and for clinically important subgroups, using ninefold cross validation
- List of variables selected and their associated parameter estimates
- Estimated odds ratios

The purpose of this information was to determine whether there were compelling statistical differences between SLS levels to support one specific choice, which in most cases we did not find. Then, surgeons examined both the full model and the variables retained at each successively more parsimonious (smaller p values) level of SLS. The goal was to choose the SLS that produced the most parsimonious model while not eliminating any variables that were thought to be important for face validity. In some instances, variables might be retained in several models with different SLS p values, but their specific coding was slightly different (eg, diabetes requiring insulin control as a separate variable, versus diabetes requiring either oral agents or insulin).

The following examples demonstrate the decision-making process used to determine the final models:

1. CABG—renal failure: SLS $p = 0.1$ selected because $p = 0.05$ resulted in loss of angiotensin-converting enzyme/angiotensin-receptor blocker in urgent/emergent patients, 2b/3a platelet inhibitors, and preoperative inotropic support variables
2. CABG—prolonged ventilation: SLS $p = 0.1$ selected because SLS $p = 0.05$ would have resulted in loss of recent cerebrovascular accident, history of mediastinal radiation, and coexisting severe mitral or aortic insufficiency variables
3. Valve procedures—stroke: SLS $p = 0.1$ selected because SLS $p = 0.05$ resulted in loss of hematocrit as a predictor variable
4. Valve procedures—prolonged ventilation: SLS $p = 0.1$ selected because SLS $p = 0.05$ resulted in omission of mitral stenosis variable
5. Valve procedures—renal failure in valve models: SLS $p = 0.05$ selected because SLS $p = 0.01$ resulted in loss

of emergent/emergent salvage, moderate/severe aortic insufficiency, severe tricuspid insufficiency, and New York Heart Association class IV heart failure variables

In some instances, practical considerations overweighed what clinicians might have preferred. For example, for the combined DSWI risk model, the full model ($SLS p = 1$) contained 63 patient risk factors and 222 parameters. That was clearly too many for the number of available endpoints and would almost certainly lead to nonconvergence or “noisy” models. It was therefore necessary to select a model with $SLS p = 0.1$, which reduced the number of variables and parameters to 25 and 63, respectively. In doing so, however, some clinically relevant risk factors were lost from the model, including mediastinal radiation, steroid use (although immunosuppression was retained), home oxygen, and liver disease.

Further detailed descriptions of the statistical approaches used in developing these models, including discrimination and calibration, are provided in Part 2 of this report.

Conclusion

Comprehensive new STS risk models have been developed based on the most contemporary data available, with the primary goal of optimizing predictive accuracy for case-mix adjustment. A structured, standardized approach to model development considered previous STS models, missing data percentages, available sample size, number of endpoints, and association of variables with endpoints. Final model development used backward selection to estimate the most parsimonious model that retained clinically essential risk factors and achieved the desired statistical performance. This strategy provided the highest likelihood of model convergence in production mode and generalizability to future data. “Big data” and machine learning approaches may some day further advance the science of risk modeling by incorporating heretofore unrecognized patterns and associations.

References

1. O'Brien SM, Feng L, He X, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery risk models: part 2—statistical methods and results. *Ann Thorac Surg* 2018;105:1419–28.
2. Kouchoukos NT, Ebert PA, Grover FL, Lindesmith GG. Report of the ad hoc Committee on Risk Factors for Coronary Artery Bypass Surgery. *Ann Thorac Surg* 1988;45:348–9.
3. Shahian DM, Jacobs JP, Badhwar V, D'Agostino RS, Bavaria JE, Prager RL. Risk aversion and public reporting. Part 1: observations from cardiac surgery and interventional cardiology. *Ann Thorac Surg* 2017;104:2093–101.
4. Shahian DM, Jacobs JP, Badhwar V, D'Agostino RS, Bavaria JE, Prager RL. Risk aversion and public reporting. Part 2: mitigation strategies. *Ann Thorac Surg* 2017;104:2102–10.
5. Shahian DM, Edwards FH, Ferraris VA, et al. Quality measurement in adult cardiac surgery: part 1—conceptual framework and measure selection. *Ann Thorac Surg* 2007;83(4 Suppl):S3–12.
6. O'Brien SM, Shahian DM, DeLong ER, et al. Quality measurement in adult cardiac surgery: part 2—statistical

- considerations in composite measure scoring and provider rating. *Ann Thorac Surg* 2007;83(4 Suppl):S13-26.
7. Rankin JS, Badhwar V, He X, et al. The Society of Thoracic Surgeons mitral valve repair/replacement plus coronary artery bypass grafting composite score: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. *Ann Thorac Surg* 2017;103:1475-81.
 8. Shahian DM, He X, Jacobs JP, et al. The Society of Thoracic Surgeons composite measure of individual surgeon performance for adult cardiac surgery: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. *Ann Thorac Surg* 2015;100:1315-25.
 9. Badhwar V, Rankin JS, He X, et al. The Society of Thoracic Surgeons mitral repair/replacement composite score: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. *Ann Thorac Surg* 2016;101:2265-71.
 10. Shahian DM, He X, Jacobs JP, et al. The STS AVR+CABG composite score: a report of the STS Quality Measurement Task Force. *Ann Thorac Surg* 2014;97:1604-9.
 11. Shahian DM, He X, Jacobs JP, et al. The Society of Thoracic Surgeons isolated aortic valve replacement (AVR) composite score: a report of the STS Quality Measurement Task Force. *Ann Thorac Surg* 2012;94:2166-71.
 12. Clark RE. It is time for a national cardiothoracic surgical data base. *Ann Thorac Surg* 1989;48:755-6.
 13. Edwards FH, Clark RE, Schwartz M. Coronary artery bypass grafting: The Society of Thoracic Surgeons National Database experience. *Ann Thorac Surg* 1994;57:12-9.
 14. Edwards FH, Grover FL, Shroyer AL, Schwartz M, Bero J. The Society of Thoracic Surgeons National Cardiac Surgery Database: current risk assessment. *Ann Thorac Surg* 1997;63:903-8.
 15. Shroyer AL, Grover FL, Edwards FH. 1995 coronary artery bypass risk model: The Society of Thoracic Surgeons Adult Cardiac National Database. *Ann Thorac Surg* 1998;65:879-84.
 16. Shroyer AL, Plomondon ME, Grover FL, Edwards FH. The 1996 coronary artery bypass risk model: The Society of Thoracic Surgeons Adult Cardiac National Database. *Ann Thorac Surg* 1999;67:1205-8.
 17. Edwards FH, Peterson ED, Coombs LP, et al. Prediction of operative mortality after valve replacement surgery. *J Am Coll Cardiol* 2001;37:885-92.
 18. Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *Ann Thorac Surg* 2003;75:1856-64.
 19. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1—coronary artery bypass grafting surgery. *Ann Thorac Surg* 2009;88(1 Suppl):S2-22.
 20. O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. *Ann Thorac Surg* 2009;88(1 Suppl):S23-42.
 21. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3—valve plus coronary artery bypass grafting surgery. *Ann Thorac Surg* 2009;88(1 Suppl):S43-62.
 22. Shahian DM, Edwards FH, Jacobs JP, et al. Public reporting of cardiac surgery performance: part 1—history, rationale, consequences. *Ann Thorac Surg* 2011;92(3 Suppl):S2-11.
 23. Shahian DM, Edwards FH, Jacobs JP, et al. Public reporting of cardiac surgery performance: part 2—implementation. *Ann Thorac Surg* 2011;92(3 Suppl):S12-23.
 24. Shahian DM, Normand SL. Comparison of "risk-adjusted" hospital outcomes. *Circulation* 2008;117:1955-63.
 25. Shahian DM, He X, Jacobs JP, et al. Issues in quality measurement: target population, risk adjustment, and ratings. *Ann Thorac Surg* 2013;96:718-26.
 26. National Academies of Sciences, Engineering, and Medicine. *Accounting for Social Risk Factors in Medicare Payment*. Washington, DC: National Academies Press; 2017.
 27. Shahian DM, He X, O'Brien SM, et al. Development of a clinical registry-based 30-day readmission measure for coronary artery bypass grafting surgery. *Circulation* 2014;130:399-409.
 28. Jones RH, Hannan EL, Hammermeister KE, et al. Identification of preoperative variables needed for risk adjustment of short-term mortality after coronary artery bypass graft surgery. The Working Group Panel on the Cooperative CABG Database Project. *J Am Coll Cardiol* 1996;28:1478-87.
 29. Tu JV, Sykora K, Naylor CD. Assessing the outcomes of coronary artery bypass graft surgery: how many risk factors are enough? Steering Committee of the Cardiac Care Network of Ontario. *J Am Coll Cardiol* 1997;30:1317-23.
 30. Harrell FE, Jr. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. New York: Springer; 2015.
 31. Afilalo J, Mottillo S, Eisenberg MJ, et al. Addition of frailty and disability to cardiac surgery risk scores identifies elderly patients at high risk of mortality or major morbidity. *Circ Cardiovasc Qual Outcomes* 2012;5:222-8.
 32. Afilalo J, Eisenberg MJ, Morin JF, et al. Gait speed as an incremental predictor of mortality and major morbidity in elderly patients undergoing cardiac surgery. *J Am Coll Cardiol* 2010;56:1668-76.
 33. Adams JL. *The Reliability of Provider Profiling: A Tutorial*. RAND Health, prepared for the National Committee for Quality Assurance. Santa Monica, CA: RAND Corporation; 2009.
 34. Gorodeski EZ, Starling RC, Blackstone EH. Are all readmissions bad readmissions? *N Engl J Med* 2010;363:297-8.

The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2—Statistical Methods and Results



Sean M. O'Brien, PhD, Liqi Feng, MS, Xia He, MS, Ying Xian, MD, PhD, Jeffrey P. Jacobs, MD, Vinay Badhwar, MD, Paul A. Kurlansky, MD, Anthony P. Furnary, MD, Joseph C. Cleveland, Jr, MD, Kevin W. Lobdell, MD, Christina Vassileva, MD, Moritz C. Wyler von Ballmoos, MD, PhD, Vinod H. Thourani, MD, J. Scott Rankin, MD, James R. Edgerton, MD, Richard S. D'Agostino, MD, Nimesh D. Desai, MD, PhD, Fred H. Edwards, MD, and David M. Shahian, MD

Duke Clinical Research Institute, Duke University Medical Center, Durham, North Carolina (SMO, LF, XH, YX); Division of Cardiac Surgery, Johns Hopkins University School of Medicine, Baltimore, Maryland (JPJ); Division of Cardiovascular Surgery, Johns Hopkins All Children's Heart Institute, St. Petersburg, Florida (JPJ); Department of Cardiovascular and Thoracic Surgery, West Virginia University, Morgantown, West Virginia (VB, JSR); Division of Cardiac Surgery, Columbia University, New York, New York (PAK); Starr-Wood Cardiothoracic Group, Portland, Oregon (APF); Division of Cardiothoracic Surgery, University of Colorado Anschutz School of Medicine, Aurora, Colorado (JCC); Atrium Health, Cardiovascular and Thoracic Surgery, Charlotte, North Carolina (KWL); Division of Cardiac Surgery, University of Massachusetts Medical School, Worcester, Massachusetts (CV); Houston Methodist DeBakey Heart and Vascular Center, Houston, Texas (MCWvB); Department of Cardiac Surgery, MedStar Heart and Vascular Institute, Georgetown University, Washington, DC (VHT); The Heart Hospital Baylor Plano, Plano, Texas (JRE); Division of Thoracic and Cardiovascular Surgery, Lahey Hospital and Medical Center, Burlington, Massachusetts (RSD); Division of Cardiothoracic Surgery, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania (NDD); Department of Surgery, University of Florida, Gainesville, Florida (FHE); and Department of Surgery and Center for Quality and Safety, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts (DMS)

Background. The Society of Thoracic Surgeons (STS) uses statistical models to create risk-adjusted performance metrics for Adult Cardiac Surgery Database (ACSD) participants. Because of temporal changes in patient characteristics and outcomes, evolution of surgical practice, and additional risk factors available in recent ACSD versions, completely new risk models have been developed.

Methods. Using July 2011 to June 2014 ACSD data, risk models were developed for operative mortality, stroke, renal failure, prolonged ventilation, mediastinitis/deep sternal wound infection, reoperation, major morbidity or mortality composite, prolonged postoperative length of stay, and short postoperative length of stay among patients who underwent isolated coronary artery bypass grafting surgery ($n = 439,092$), aortic or mitral valve surgery ($n = 150,150$), or combined valve plus coronary artery bypass grafting surgery ($n = 81,588$). Separate models were developed for each procedure and endpoint except mediastinitis/deep sternal wound infection, which was analyzed in a combined model because of its

infrequency. A surgeon panel selected predictors by assessing model performance and clinical face validity of full and progressively more parsimonious models. The ACSD data (July 2014 to December 2016) were used to assess model calibration and to compare discrimination with previous STS risk models.

Results. Calibration in the validation sample was excellent for all models except mediastinitis/deep sternal wound infection, which slightly underestimated risk and will be recalibrated in feedback reports. The c-indices of new models exceeded those of the last published STS models for all populations and endpoints except stroke in valve patients.

Conclusions. New STS ACSD risk models have generally excellent calibration and discrimination and are well suited for risk adjustment of STS performance metrics.

(Ann Thorac Surg 2018;105:1419–28)

© 2018 by The Society of Thoracic Surgeons

Accepted for publication March 9, 2018.

The STS Executive Committee approved this document.

Address correspondence to Dr Shahian, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114; email: dshahian@partners.org.

The [Supplemental Material](#) can be viewed in the online version of this article [<https://doi.org/10.1016/j.athoracsur.2018.03.003>] on <http://www.annalsthoracicsurgery.org>.

Abbreviations and Acronyms

ACSD	= Adult Cardiac Surgery Database
AVR	= aortic valve replacement
BMI	= body mass index
BSA	= body surface area
CABG	= coronary artery bypass grafting surgery
DSWI	= deep sternal wound infection/mediastinitis
MVR	= mitral valve replacement
MVr	= mitral valve repair
PLOS	= postoperative length of stay
STS	= The Society of Thoracic Surgeons

Risk models are used for multiple purposes in adult cardiac surgery including quality measurement, clinical practice improvement, voluntary public reporting, and research. These risk models are used by The Society of Thoracic Surgeons (STS) to benchmark participant outcomes in comparison with national aggregate data in STS feedback reports; to enable case-mix adjustment in the calculation of participant and individual surgeon composite performance measures; and to support the STS voluntary public reporting initiative. To maximize the validity of its performance metrics, the STS has developed a portfolio of risk models that are customized for specific procedure populations and that adjust for numerous patient preoperative factors.

The 2008 STS adult cardiac surgery risk models were based on data from 2002 to 2006 [1–3]. Since the publication of these models, a recalibration factor has been applied to each subsequent harvest period so that the ratio of observed to expected outcomes would equal 1. In the decade since publication of the 2008 risk models, successive Adult Cardiac Surgery Database (ACSD) versions have been introduced to account for temporal changes in procedures, patient populations, surgical practices, outcomes, and the identification of new risk factors. Using the most recent data version available at the time of this analysis, we sought to develop a completely new set of STS adult cardiac surgery risk models.

Part 1 of this report [4] provides a detailed background and conceptual framework for the risk model update and provides a high-level methodologic summary of the update process. In Part 2, we provide the detailed statistical methods and results.

Patients and Methods

Endpoints

Risk models were developed for the following nine endpoints chosen for consistency with prior STS risk models and current performance metrics (eg, STS composite scores): (1) operative mortality, defined in all STS databases as all deaths, regardless of cause, occurring during the hospitalization in which the operation was performed

even if after 30 days (includes patients transferred to other acute care facilities), and all deaths, regardless of cause, occurring after discharge from the hospital but before the end of the 30th postoperative day; (2) stroke—an acute episode of focal or global neurologic dysfunction caused by brain, spinal cord, or retinal vascular injury as a result of hemorrhage or infarction in which the neurologic dysfunction lasts for more than 24 hours; (3) renal failure—a new requirement for dialysis or meeting the RIFLE (Risk, Injury, Failure, Loss of kidney function, and End-stage kidney disease) criteria based on creatinine levels or glomerular filtration rate [5]; (4) prolonged ventilation or reintubation—more than 24 hours; (5) mediastinitis/deep sternal wound infection (DSWI) occurring during the index hospitalization or within 30 days of operation; (6) reoperation for bleeding, tamponade, or any cardiac reason; (7) major morbidity or mortality—a composite defined as the occurrence of any one or more of the above endpoints; (8) prolonged postoperative length of stay (PLOS)—PLOS more than 14 days (alive or dead); and (9) short PLOS, defined as PLOS less than 6 days and patient alive at discharge. The follow-up period for endpoint definitions was from operation until the latter of hospital discharge or 30 days for mortality and DSWI and until hospital discharge for all other endpoints.

Endpoints with notable definition changes compared with the STS 2008 models included stroke (changed duration of symptoms from more than 72 hours to more than 24 hours), reoperation (changed from any reason to any cardiac reason), DSWI (added mediastinitis and included both in-hospital and 30-day timeframe), and renal failure (definition changed to more closely align with RIFLE criteria [5]).

Study Cohort

Models were developed and evaluated using data from July 1, 2011, to December 31, 2016, and were limited to the three major procedure populations that have been designated for outcomes reporting in the STS participant feedback report: (1) isolated CABG; (2) isolated valve; and (3) valve plus CABG. Data collected under STS version 2.73 (July 1, 2011, to June 30, 2014) were used to develop the models and perform a preliminary internal assessment of discrimination and calibration. Data collected under STS version 2.81 (July 1, 2014, to December 31, 2016) were used to assess model performance in a separate patient sample.

The valve cohort includes operations for aortic valve replacement (AVR), mitral valve replacement (MVR), and mitral valve repair (MVr). The valve plus CABG population includes AVR plus CABG, MVR plus CABG, and MVr plus CABG. Definitions of these populations are provided in the [Supplemental Material](#). Briefly, each operation type includes patients undergoing a stand-alone operation and excludes planned major concomitant operations with a few exceptions, most notably, that concomitant tricuspid valve repair, surgical ablation for atrial fibrillation, or repair of atrial septal defect are allowed concomitantly with MVR or MVr in the valve and

valve plus CABG populations. Patients on dialysis preoperatively were excluded from models predicting new-onset postoperative renal failure.

Among 1,556,593 records for patients aged 18 to 110 years undergoing a cardiac operation at a participating site in the United States or Canada during the study period, 1,250,165 (80%) records met criteria for one of the three major procedure populations (Table 1) and were included in the development cohort ($n = 670,830$) or validation cohort (579,335).

Risk Models

For each endpoint except DSWI, separate risk models were developed for each major procedure population (8 endpoints \times 3 populations = 24 risk models). For DSWI, the low number of endpoint events in the valve ($n = 244$) and valve plus CABG ($n = 285$) populations prompted concern that models in these populations may prove to be inaccurate because of overfitting the data. To mitigate overfitting, we developed a single DSWI model combining all three procedure populations. The DSWI models used indicator variables to adjust for operation type (eg, AVR, MVr, MVR, and so forth) and included interaction terms to account for the importance of selected risk factors that differ across these operation types.

Selection of Candidate Predictor Variables

The 2018 STS risk models were developed using data from version 2.73, but these models will be applied to patients entered into the STS ACSD using versions 2.81 and later. Accordingly, to be an acceptable candidate variable, it was necessary to assure that the variable was present in version 2.73 and in version 2.81 (or a similar, mappable analogue in the latter). Because the main goal of the models is to adjust for case mix, only preprocedural patient variables were considered for inclusion.

To begin the selection process, each surgeon member of the working group ($n = 10$) independently reviewed a list of 187 potentially relevant preprocedure factors from

the v2.73 data collection form and used an online questionnaire to rate his or her a priori assessment of each variable's prognostic potential. Variables identified as potential risk factors by at least four of the 10 surgeons were retained for further consideration and were discussed in detail in a series of conference calls. To facilitate this discussion, each variable's frequency distribution and percentage of missing data were tabulated overall and across operation types.

Missing data frequency was less than 1% for the majority of preprocedural variables. Those few variables with missing data rates greater than 5%, or variables associated with a test or study that had not been performed in more than 5% of the relevant study population, were also excluded. Specific examples of excluded variables are described in Part 1 of this report [4].

Considerations regarding adjustment of outcomes measures for socioeconomic status or sociodemographic factors (eg, race, ethnicity, education, income, payer [eg, Medicare-Medicaid dual eligible status]) are discussed in detail in Part 1 of this report [4]. In general, we based our modeling decisions on principles from epidemiology and causal inference, evaluating those available socioeconomic status or sociodemographic risk factors potentially having an empirical association with outcomes and relevant to case-mix adjustment, while avoiding more philosophical considerations.

Surgery date was included as a candidate predictor to adjust for temporal trends in endpoint occurrence rates and detection rates across the 3-year development period. Risk calculators implementing these models will account for time trends by predicting risk standardized to a January 1, 2014, surgery date.

Simulations to Assess Statistical Precision and Overfitting

When the number of predictors in a model is too large in relation to the available sample size, the estimated numerical coefficients are likely to be inaccurate because of overfitting the current study data [6]. Using a data-driven variable selection procedure can reduce the number of predictors in a model but may not mitigate overfitting because each predictor tested for inclusion in the model has the potential to be selected because of overfitting [7].

To assess the potential statistical accuracy of risk models based on this project's available sample size and candidate risk factors, a simulation study was conducted. This involved creating 200 bootstrap samples by sampling records with replacement from the overall development sample and using each bootstrap sample to estimate two models for each combination of model population and endpoint, based on the surgeon panel's proposed list of candidate predictors. The first model included the entire set of proposed candidate predictor variables, a so-called full model. The second model started with the same set of predictors but applied backward selection with a significance threshold of 0.05. Regression coefficients from each bootstrap sample were then used to calculate predicted risk estimates for each patient in the overall development sample. Ideally, in a setting of high statistical precision,

Table 1. Sample Sizes for Model Development and Evaluation

Procedure	Development ^a	Validation ^b
Overall	670,830	579,335
CABG	439,092	385,179
Valve	150,150	129,511
AVR	87,629	72,719
MVR	26,850	25,888
MVr	35,671	30,904
Valve+CABG	81,588	64,645
AVR+CABG	55,064	43,822
MVR+CABG	9,227	8,737
MVr+CABG	17,297	12,086

^a July 2011 to June 2014. ^b July 2014 to December 2016.

AVR = aortic valve replacement; CABG = coronary artery bypass grafting surgery; MVR = mitral valve replacement; MVr = mitral valve repair.

the predicted risk estimate for the same patient and endpoint should not vary depending on which bootstrap sample was used to estimate regression coefficients.

To quantify estimation error, we estimated the average Pearson correlation between risk estimates for the same patient across all possible pairs of bootstrap samples. This was done separately for each combination of population (CABG, valve, valve plus CABG), endpoint, and modeling strategy (all predictors, backward selection). Results indicated that predicted risk estimates were generally stable with Pearson correlation coefficients greater than 0.90 for most population and endpoints combinations whether retaining all predictors or using backward selection.

Models with low consistency across bootstrap samples were those with relatively few endpoint events, including stroke in the valve and valve plus CABG populations (correlations 0.58 to 0.69) and DSWI in the valve and valve plus CABG populations (correlations 0.25 to 0.41). These findings led the model committee to consider both predictive accuracy and parsimony in the final selection of candidate predictors and to estimate a single combined model for DSWI, instead of separate DSWI models for each procedure population, as mentioned previously.

Optimal Coding of Candidate Covariates

We attempted to achieve the most computationally efficient and clinically relevant coding, or parameterization, of candidate variables. That typically involved collapsing or combining clinically related or collinear variables, and was particularly important in cases where multiple STS variables relate to a single underlying clinical concept, for example, insurance status (12 variables), previous cardiac interventions (31 STS variables), and preoperative arrhythmias (7 STS variables). In some instances, uncommon but important variables were combined with other related variables (eg, catheter-based assist devices and extracorporeal membrane oxygenation were combined with shock, their usual indication).

For some variables, informal exploratory analyses using STS data from an earlier period (2007 to 2011) helped to determine the optimal modeling strategy. For example, we used data from 2007 to 2011 to explore how best to model body mass index (BMI) and body surface area (BSA) given that both variables describe aspects of a patient's body habitus and are highly correlated. For these investigations, we initially estimated a multivariable model for mortality that did not adjust for BSA or BMI but included all other preoperative factors from the published STS 2008 mortality model. After fitting this model to data from 2007 to 2011, we then compared observed versus predicted mortality rates across subgroups defined by categorizations of BSA and BMI as well as sex. The observed pattern of residuals indicated that BSA and BMI were both independently associated with mortality and that inclusion of both variables was needed to capture variation in the residuals. When the model was reestimated after including BSA but not BMI, the pattern of residuals indicated a U-shape relationship between BMI and mortality, leading to the inclusion of both linear and

quadratic terms for BMI. We investigated the following approaches for BSA and BMI: BSA linear; BSA quadratic; interaction between BSA and sex; BMI linear; BMI quadratic; and BMI alternatives. In some instances, extreme values were truncated (eg, BMI values greater than 50 were mapped to 50).

Similar analyses were conducted to explore modeling issues for insurance status, race, myocardial infarction history, and history of prior procedures, and to explore the functional form of various other continuous variables.

Selection of Final Covariates

After choosing the list of candidate covariates, the final set of covariates for each model were selected. For each combination of population and endpoint, we estimated a full model that included all candidate covariates and a set of reduced models that were chosen by backward variable selection. To estimate the optimal significance level for backward selection, we repeated the backward selection process using five different significance levels (0.0001, 0.001, 0.01, 0.05, and 0.1) and estimated performance metrics for the resulting models. The goal of this analysis was to select the optimum significance level to use for each combination of population and endpoint. Because of overfitting the data, model performance is likely to be overestimated when models are developed and naively tested in the same sample of data. To obtain approximately unbiased performance estimates, each full model and the backward selection process was repeated in 200 bootstrap samples drawn with replacement from the original development sample.

To assess overfitting, we applied estimated regression coefficients from the bootstrap sample to patients in the overall development sample and then entered each patient's calculated risk score (log-odds) into a univariable logistic regression model predicting the endpoint. The slope coefficient for the risk score in this model was interpreted as a measure of overfitting, with a slope of 1.0 indicating perfect calibration and a slope less than 1 indicating possible overfitting [6]. Discrimination was assessed by calculating the c-statistic (the area under the receiver-operating characteristics curve) and using a bootstrap adjustment to correct for optimism [7].

To assess calibration, the backward selection process was subsequently repeated using ninefold cross validation. For each cross-validation replicate, models were developed in an 8/9 training sample and evaluated for calibration in a 1/9 testing sample. Calibration was assessed graphically by plotting observed versus expected endpoint event rates across deciles of predicted risk among patients in each testing sample. That was done for the full model and for each significance level when using backward selection. The main objective of this exercise was to determine whether there were compelling statistical differences between significance levels to support one particular choice.

In the absence of compelling statistical differences between the performance of various models, the final model was chosen by surgeon members of the working group, as

described in Part 1 of this report [4]. Beginning with the full model, surgeons carefully reviewed the predictors in each model (full, and using backward selection criteria $p = 0.1, 0.05, 0.01, 0.001$, and 0.0001). Each progressively more parsimonious model was evaluated to be certain that no variables had been eliminated that would jeopardize clinical face validity. Generally, the most statistically parsimonious model that did not compromise clinical face validity was chosen as the final model.

Missing Data and Imputation Strategies

Covariate data were missing in fewer than 5% of cases in each procedure population for all but one candidate covariate (aortic root abscess in AVR and AVR plus CABG; missing = 13%). Overall, 15% of records had missing or unknown mortality data for at least one component of the operative mortality definition. Rates of missing or unknown data were 0.06% for discharge mortality status and 15.0% for 30-day mortality status. Previous linkage of the STS ACSD to the Social Security Death Master File [8] reveals that capture of 30-day deaths occurring before discharge is highly accurate, and that these in-hospital deaths represent the majority (79%) of all 30-day deaths. Capture of the remaining 30-day deaths occurring after discharge was less complete and warranted improvement. Consequently, in 2016, the STS implemented more stringent requirements for all data fields related to operative mortality. As of January 1, 2016, participants were not included in the benchmark population for STS performance metrics, nor were these participants eligible to receive an STS star rating unless their rate of missing data for 30-day mortality and discharge mortality was less than 5% missing or unknown; in January 2017 this threshold was further decreased to 2%.

Missing data rates for endpoints other than mortality were less than 0.25%. For initial exploratory and variable

selection analyses, missing covariate and endpoint values were handled using a simple single imputation strategy. Values were imputed to the most common category of binary or categorical variables and to the median or subgroup-specific median of continuous variables. This single imputation strategy was previously validated for the 2008 STS risk models by demonstrating that coefficients and predicted risk estimates obtained using single imputation were similar to the gold standard of multiple imputation [1].

After finalizing the selection of model covariates, as described above, regression coefficients were subsequently reestimated using a multiple imputation strategy for covariates with more than 5% missing data and for all endpoints. The principle motivation for using multiple imputation was to make efficient use of data from the discharge mortality status field when imputing operative mortality status among patients who were discharged alive. Multiple imputation was implemented using the method of chained equations as implemented in the SAS software (SAS Institute, Cary, NC) PROC MI procedure with the full conditional specification option [9, 10]. To avoid bias due to perfect prediction [11], separate imputation models were estimated for discharge deaths and discharge survivors. To speed computation and resolve convergence errors, covariates with less than 5% missing data were imputed by single imputation before estimating the multiple imputation model.

Final Model Assessment

The validation sample was created by applying the study's inclusion criteria to STS data for the period July 1, 2014, to December 31, 2016, as the goal was to assess model performance in future data. Data from hospitals with more than 5% missing data for an endpoint within a procedure population were excluded from validation analyses for that population and endpoint. Discrimination was quantified

Table 2. Percentage and Number of Endpoint Events by Model Population in Development Sample

Endpoint Events	All (n = 670,830)	CABG (n = 439,092)	Valve (n = 150,150)	Valve + CABG (n = 81,588)
Operative mortality	2.9% 16,792/569,998	2.4% 8,852/373,683	3.2% 4,004/126,204	5.6% 3,936/70,111
Stroke	1.5% 9,866/669,561	1.3% 5,621/438,385	1.5% 2,237/149,800	2.5% 2,008/81,376
Renal failure	2.7% 17,202/648,808	2.2% 9,381/424,888	2.7% 3,868/145,454	5.0% 3,953/78,466
Prolonged ventilation	10.9% 72,984/670,830	9.3% 40,974/439,092	11.1% 16,604/150,150	18.9% 15,406/81,588
Reoperation	3.1% 20,872/670,778	2.4% 10,327/439,060	4.2% 6,371/150,137	5.1% 4,174/81,581
Composite morbidity and mortality	17.4% 101,180/581,976	15.0% 56,984/380,491	18.4% 23,724/129,140	28.3% 20,472/72,345
Prolonged PLOS	6.6% 44,533/670,428	5.0% 22,091/438,867	8.0% 11,941/150,024	12.9% 10,501/81,537
Short PLOS	42.7% 286,362/670,428	48.3% 211,820/438,867	37.4% 56,130/150,024	22.6% 18,412/81,537
DSWI	0.3% 1,875/669,392	0.3% 1,346/438,270	0.2% 244/149,778	0.4% 285/81,344

CABG = coronary artery bypass grafting surgery; DSWI = mediastinitis/deep sternal wound infection; PLOS = postoperative length of stay.

Table 3. Candidate Predictors

Operation type	Illicit drug use
Age	Alcohol consumption (drinks per week)
Ejection fraction	Recent pneumonia
Body mass index	Mediastinal radiation
Body surface area	Cancer diagnosis within 5 years
Sex	Diabetes/diabetes control method
Renal function (dialysis/creatinine)	Number of diseased vessels
Hematocrit	Myocardial infarction history/timing
White blood cell count	Cardiac presentation on admission
Platelet count	Race/ethnicity
ADP receptor inhibitor usage/timing of discontinuation	Status
Hypertension	ACE/ARB inhibitor within 48 hours in nonelective operation
Immunosuppressive therapy within 30 days	Heart failure class and timing
Steroids within 24 hours	Recent smoker/timing
Glycoprotein IIb/IIIa inhibitor within 24 hours	Family history of CAD
Inotropes within 48 hours	Home oxygen
Preoperative IABP	Sleep apnea
Shock/ECMO/CBA	Liver disease
PAD	Unresponsive neurologic status
Left main disease	Syncope
Proximal LAD	Previous CABG
Aortic root abscess in AVR/AVR+CABG	Previous aortic valve procedure
Mitral stenosis	Previous mitral valve procedure
Aortic stenosis	Previous transcatheter valve replacement/percutaneous valve repair
Mitral insufficiency	Previous other valve procedure
Tricuspid insufficiency	Number of previous cardiovascular surgeries
Aortic insufficiency	Previous ICD
Arrhythmia and type	PCI history/timing
Endocarditis	Previous any other cardiac intervention
Chronic lung disease	Payer/insurance type
CVD/CVA/TIA	Tricuspid valve repair performed concomitantly
Carotid stenosis	Time trend (surgery date)
Previous carotid surgery	

ACE = angiotensin-converting enzyme; ADP = adenosine diphosphate; ARB = angiotensin-receptor blocker; AVR = aortic valve replacement; CABG = coronary artery bypass grafting surgery; CAD = coronary artery disease; CBA = catheterization-based assist device; CVA = cerebrovascular accident; CVD = cardiovascular disease; ECMO = extracorporeal membrane oxygenation; IABP = intraaortic balloon pump; ICD = implantable cardioverter-defibrillator; LAD = left anterior descending artery; PAD = peripheral arterial disease; PCI = percutaneous coronary intervention; TIA = transient ischemic attack.

by the c-statistic. To provide context for interpreting discrimination results, c-statistics were calculated in the validation sample for both the current STS 2018 models and the prior STS 2008 models. Calibration was assessed by plotting observed versus expected event rates across deciles of predicted risk in the validation sample.

Results

A total of 670,830 records met study inclusion criteria and were included in the development samples for CABG (n = 439,092), valve (n = 150,150), and valve plus CABG (n = 81,588). The number of endpoint events in the development sample ranged from 1,875 for DSWI to

Table 4. C-statistics in Validation Sample for 2008 STS Risk Models and Current 2018 Risk Models

Endpoint	CABG		Valve		Valve + CABG	
	STS 2008 Models	STS 2018 Models	STS 2008 Models	STS 2018 Models	STS 2008 Models	STS 2018 Models
Operative mortality	0.791	0.804	0.760	0.775	0.753	0.761
Stroke	0.682	0.697	0.669	0.656	0.631	0.632
Renal failure	0.801	0.826	0.770	0.787	0.746	0.759
Prolonged ventilation	0.756	0.772	0.761	0.777	0.731	0.744
Reoperation	0.600	0.621	0.604	0.616	0.577	0.588
Composite morbidity and mortality	0.725	0.738	0.712	0.723	0.702	0.712
Prolonged PLOS	0.761	0.777	0.778	0.796	0.726	0.739
Short PLOS	0.707	0.716	0.723	0.732	0.716	0.726
DSWI	0.665	0.681	0.592	0.665	0.648	0.659

CABG = coronary artery bypass grafting surgery; DSWI = mediastinitis/deep sternal wound infection; PLOS = postoperative length of stay; STS = The Society of Thoracic Surgeons.

286,362 for short PLOS (Table 2). As discussed above, the relatively small number of DSWI endpoints in valve ($n = 244$) and valve plus CABG ($n = 285$) populations raised concerns about potential overfitting in these populations, and that led to a decision to estimate a single combined model for DSWI. For the other eight endpoints, the number of occurrences ranged from 2,008 for stroke in valve plus CABG to 211,820 for short PLOS in CABG.

Table 3 summarizes the final list of candidate covariates. These 65 variables were included in the "full" model for each endpoint and population and were the starting point for variable selection by backward selection with bootstrapping and cross validation, and subsequent clinical assessment by the surgeon panel. Details of how each candidate variable was parameterized in the model

are provided in the Supplemental Material. After inclusion of nonlinear, categorical, and interaction terms, the number of model parameters in the full model was 122 for CABG, 218 for valve, and 215 for valve plus CABG. The number of endpoint events per parameter in the full model ranged from 9 in the stroke model for valve plus CABG to 1,736 in the short PLOS model for CABG.

Supplemental Tables 1 to 4 in the Supplemental Material summarize risk factors in the final selected model for each population and endpoint. The number of risk factors in these models ranged from 25 in the model for stroke in valve plus CABG to 50 in the models for composite mortality or major morbidity and short PLOS in CABG. Full specifications for these models including formulas, coefficients, and intercept parameters will be publicly available from the STS website.

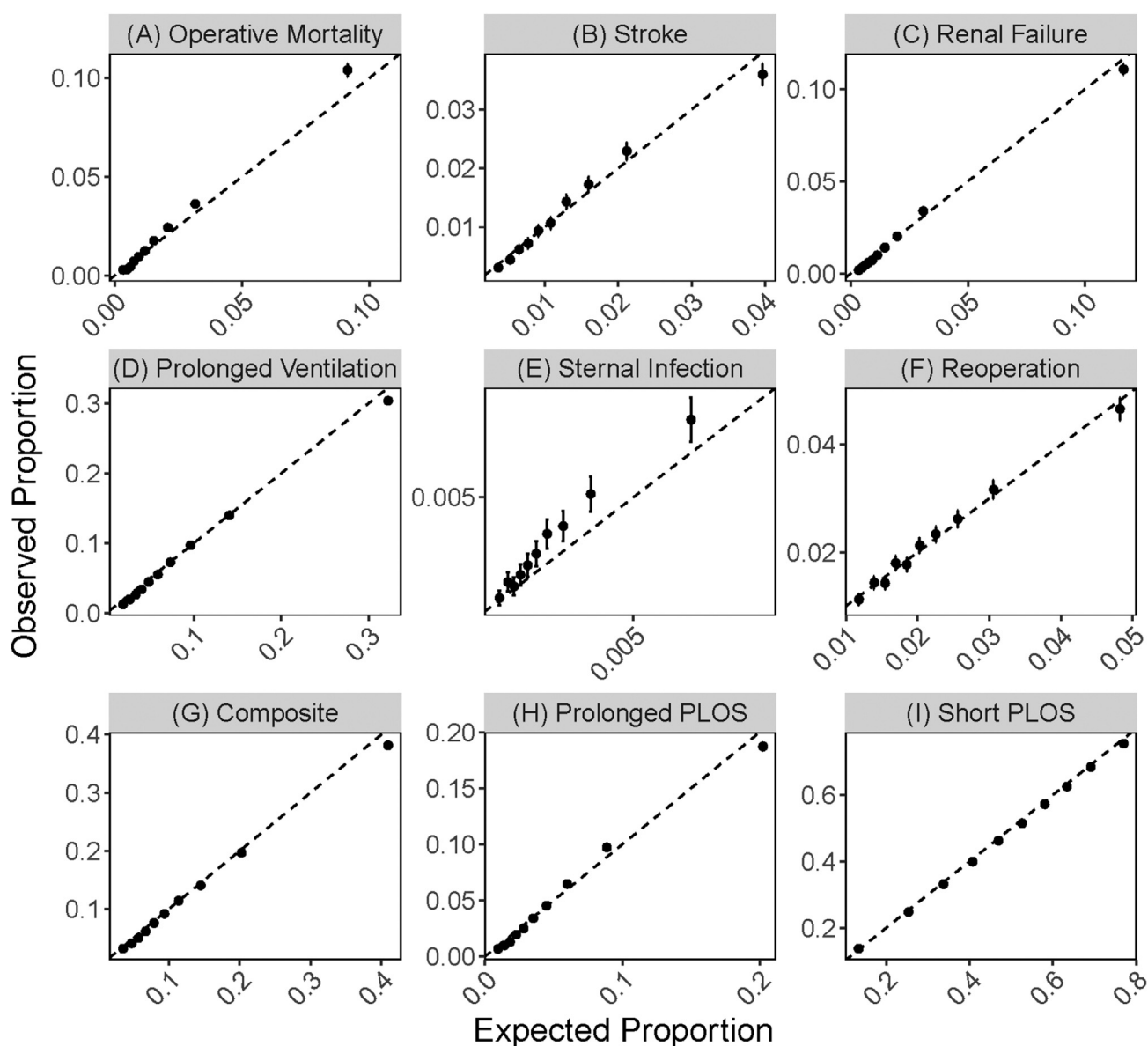


Fig 1. Calibration for major endpoints in the validation sample: coronary artery bypass grafting surgery. (PLOS = postoperative length of stay.)

Performance of the final models in the development sample was excellent for most of the population and endpoint combinations (Supplemental Material, Supplemental Tables 5, 6). Across the three model populations, the bootstrap-adjusted c-statistics were lowest for reoperation (range, 0.574 to 0.627) followed by stroke (range, 0.616 to 0.704) and were highest for renal failure (range, 0.749 to 0.810). Slopes to assess overfitting were generally close to the ideal value of 1.0 and were greater than 0.90 for all but three population-endpoint combinations. Models with slopes less than 0.90 were reoperation in valve (0.88), reoperation in valve plus CABG (0.78), and stroke in valve plus CABG (0.79). Calibration plots based on cross validation revealed acceptable calibration and no obvious violation of modeling assumptions.

After selecting the final set of models, regression coefficients were subsequently reestimated using multiple imputation to deal with missing endpoint data. After multiple imputation, the average predicted mortality risk across all populations increased from 2.50% to 2.58% (relative increase = 3%).

The c-statistics in the validation sample ranged from 0.588 for reoperation in valve plus CABG to 0.826 for renal failure in CABG. Table 4 presents c-statistics calculated in the validation sample for the final selected models and compares them with c-statistics calculated in the validation sample for the prior STS 2008 risk models. Although the DSWI model was estimated in a combined cohort that included all three procedure populations, its discrimination was assessed in each procedure population individually in Table 4. The c-statistics of the new STS models

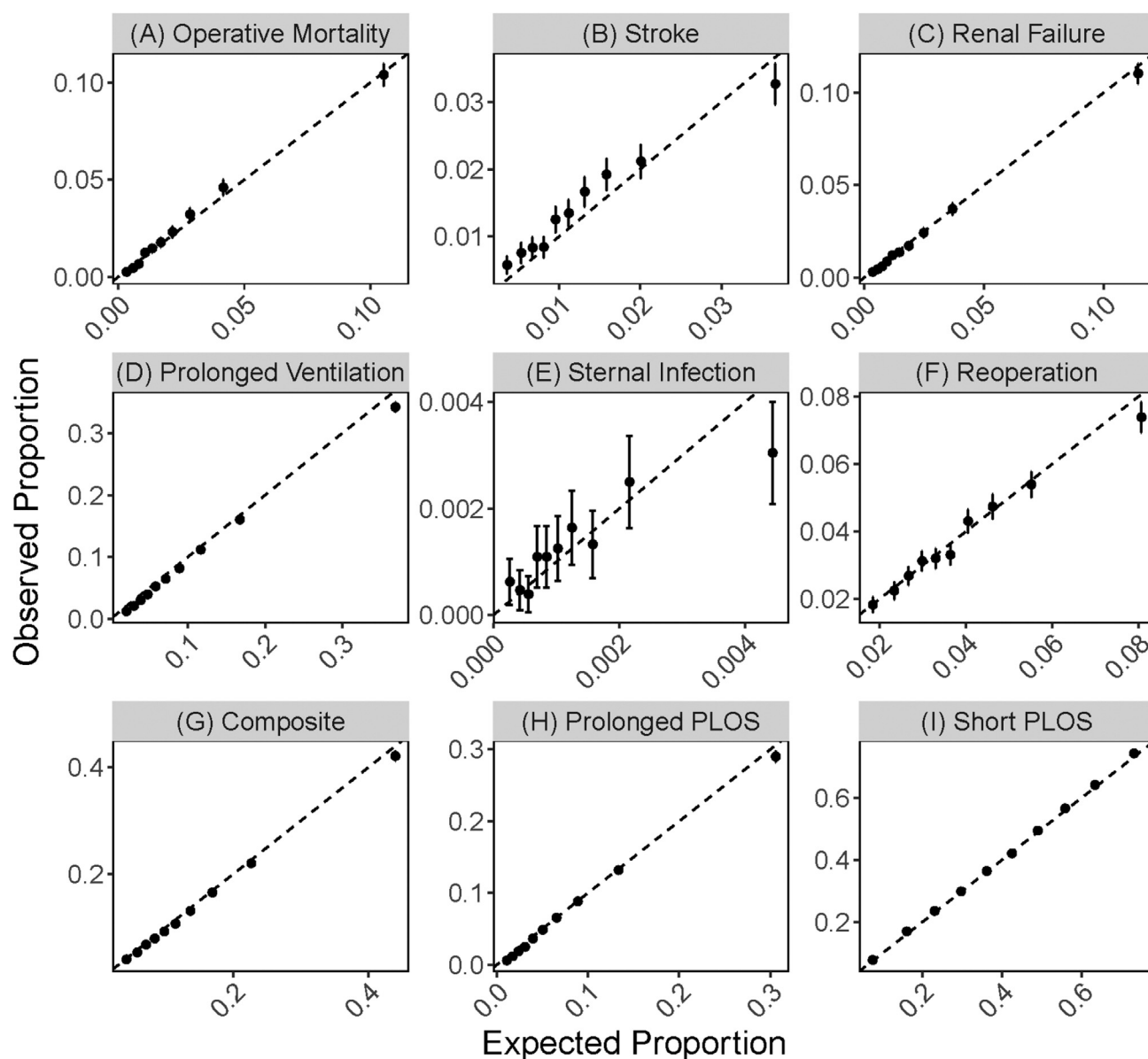


Fig 2. Calibration for major endpoints in the validation sample: valve. (PLOS = postoperative length of stay.)

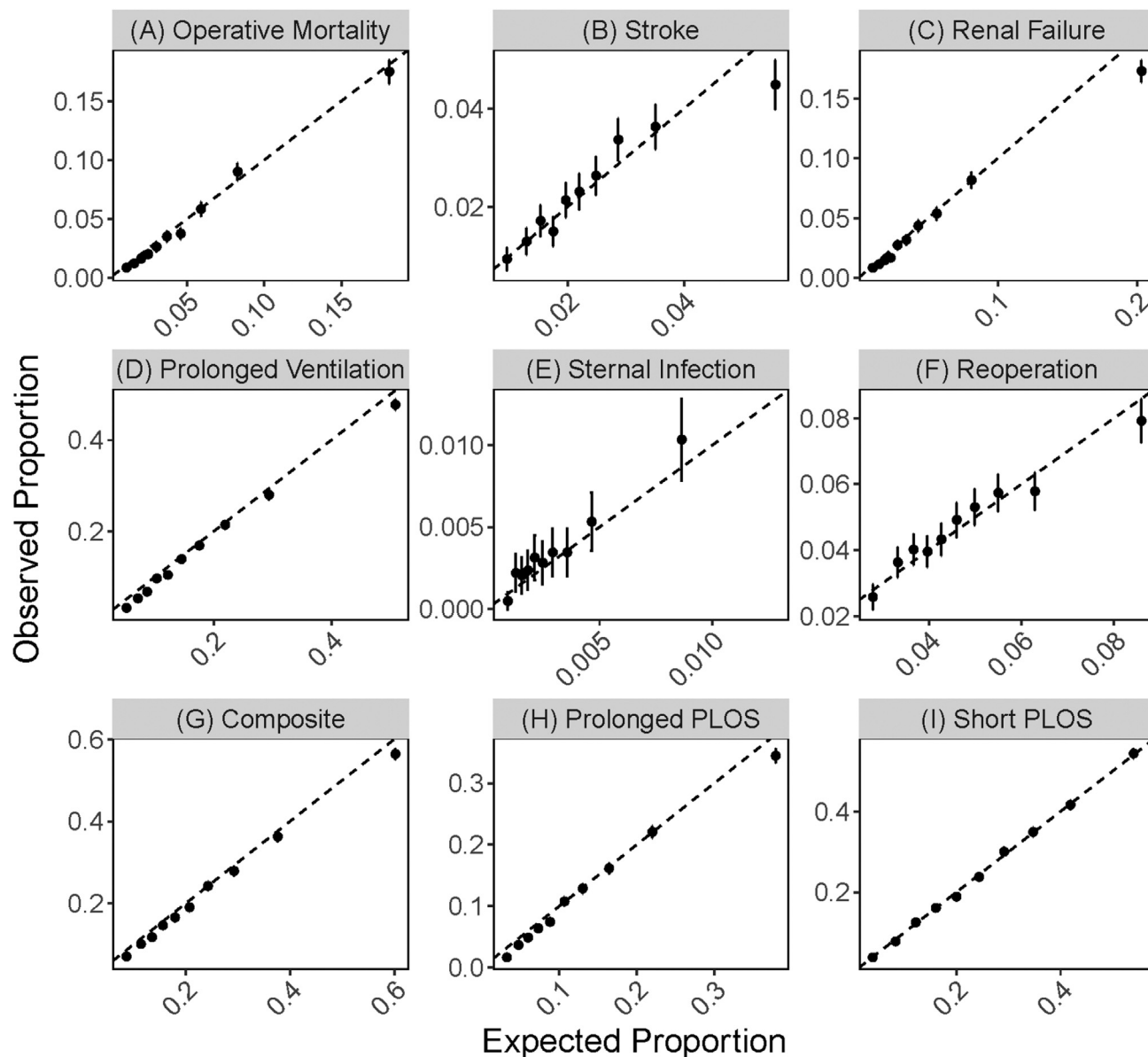


Fig 3. Calibration for major endpoints in the validation sample: valve plus coronary artery bypass grafting surgery. (PLOS = postoperative length of stay.)

exceeded those of the STS 2008 models for all populations and endpoints except for the valve model for stroke; all but two of the p values were less than 0.05 (stroke and DSWI in valve plus CABG) and most were less than 0.0001 (Supplemental Table 7).

Calibration graphs in the validation sample are presented in Figures 1, 2, and 3. These reveal excellent calibration for the vast majority of populations and endpoints. The DSWI model appears to systematically underestimate infection risk in CABG by a factor of approximately 0.80, presumably because of a somewhat higher rate of this complication in more recent data. This underestimation of risk will be corrected when these models are used to calculate observed to expected ratios in the STS feedback reports; the report methodology applies

a calibration factor that causes the expected rate to equal the observed rate within each calendar year of the reporting period. After applying the STS feedback report recalibration methodology to the validation sample, the calibration of the recalibrated DSWI model was excellent, as shown in Supplemental Figure 1. When these models are used to calculate STS composite scores, deteriorating calibration over time will be corrected automatically because model coefficients will be reestimated in the current STS data before composite scores are calculated.

Comment

We have described the development and validation of a comprehensive set of new STS adult cardiac surgical risk

models that will be used to adjust for case mix in the STS participant feedback report and the STS voluntary public reporting program. Our approach to model development incorporated several novel features including the use of simulations to assess the feasible number of predictors in relation to sample size, and the combined use of bootstrapping and cross validation to estimate model operating characteristics as a function of the significance level for variable inclusion. Because the main intended use of these models was case-mix adjustment, we did not focus on parsimony (small number of covariates) as our primary goal but rather selected the optimal covariates for accurate risk prediction using a combination of statistical and clinical face validity approaches. The models showed good calibration, and 24 of 25 models had superior discrimination compared with the STS 2008 models when evaluated in the contemporary dataset used for model validation.

References

1. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: part 1—coronary artery bypass grafting surgery. *Ann Thorac Surg* 2009;88(1 Suppl):S2–22.
2. O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: part 2—isolated valve surgery. *Ann Thorac Surg* 2009;88(1 Suppl):S23–42.
3. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 Cardiac Surgery Risk Models: part 3—valve plus coronary artery bypass grafting surgery. *Ann Thorac Surg* 2009;88(1 Suppl):S43–62.
4. Shahian DM, Jacobs JP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery risk models: part 1—background, design considerations, and model development. *Ann Thorac Surg* 2018;105:1411–8.
5. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P. Acute renal failure—definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit Care* 2004;8:R204–12.
6. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373–9.
7. Harrell FE, Jr. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. New York: Springer; 2015.
8. Jacobs JP, O'Brien SM, Shahian DM, et al. Successful linking of The Society of Thoracic Surgeons Database to Social Security data to examine the accuracy of Society of Thoracic Surgeons mortality data. *J Thorac Cardiovasc Surg* 2013;145:976–83.
9. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011;30:377–99.
10. SAS Institute Inc. *SAS/Stat 13.2 User's Guide*. Cary, NC: SAS Institute; 2014.
11. White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Comput Stat Data Anal* 2010;54:2267–75.

STS CABG COMPOSITE (0696)

ANALYSES USING 2018 DATA (1/1/2018 – 12/31/2018)

Analyses for #0696 STS CABG composite were originally performed using data from all isolated CABG operations performed July 2013 – June 2014 (n = 143,771 operations from n = 1,024 STS participants). We repeated a subset of these analyses using data from all isolated CABG operations performed in 2018 (n = 152,446 records from n = 962 participants).

Distribution of Statistical Classification Categories

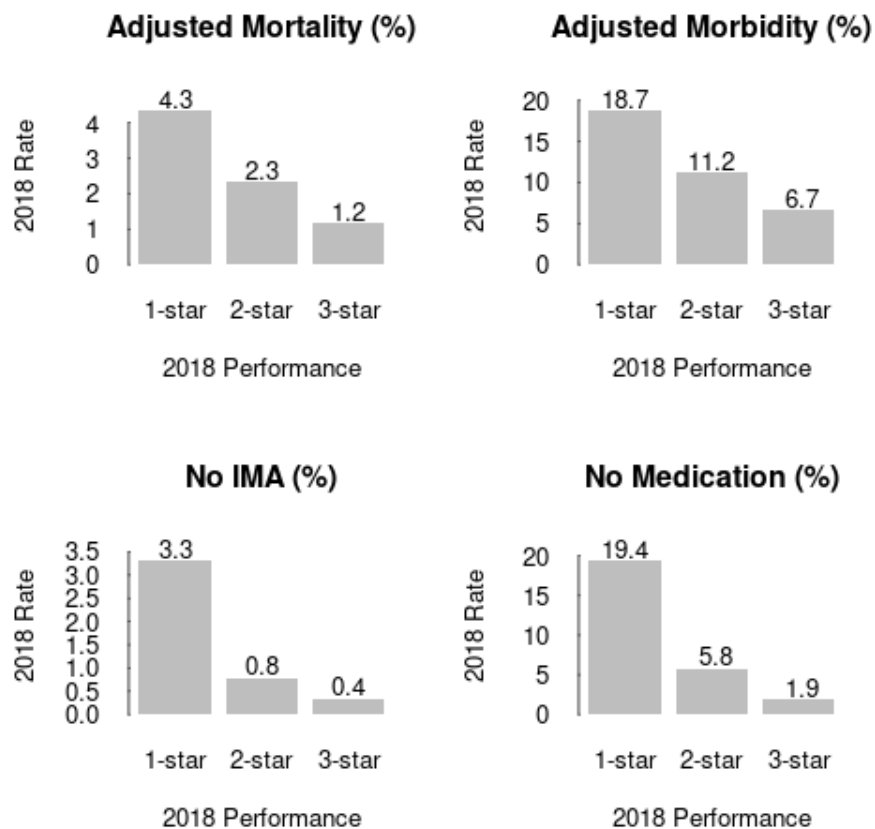
1-STAR	2-STAR	3-STAR
41 (4%)	852 (88%)	69 (7%)

Note: Assignment to 1-star or 3-star is based on 99% Bayesian probability

Signal-to-Noise Reliability (2018 data)

All Participants (n = 962)	Participants With At Least 50 Eligible Cases (n = 844)	Participants With At Least 100 Eligible Cases (n = 588)
0.64	0.66	0.67

Association of Composite Categories With Individual Quality Domain Scores



4b1 – Usability and Use

Measure 0696: Percentages for star ratings since 2010 (updated 11/2019)

	Stars	2018	2017	2016	2015	2014	2013	2012	2011	2010
0696 - STS CABG Composite	*	4.37	4.55	5.29	5.82	4.59	9.19	9	9.6	11
	**	88.27	89.21	84.65	84.4	86.64	75.86	76	76.5	75.5
	***	7.36	6.24	10	9.74	8.77	14.95	15	14	13.5