Ready-Made, Recalibrated, or Remodeled?  
Issues in the Use of Risk Indexes for Assessing Mortality After Coronary Artery Bypass Graft Surgery

Joan Ivanov, RN, MSc; Jack V. Tu, MD, PhD; C. David Naylor, MD, DPhil

Background—Risk indexes for operative mortality after cardiac surgery are used for comparative profiling of surgeons or centers. We examined whether clinicians and managers should use an existing index without modification, recalibrate it for their populations, or derive a new model altogether.

Methods and Results—Drawing on 7491 consecutive patients who underwent isolated CABG at 2 Toronto teaching hospitals between 1993 and 1996, we compared 3 strategies: (1) using a ready-made model originally derived and validated in our jurisdiction; (2) recalibrating the ready-made model to better fit the population; and (3) deriving a new model with additional risk factors. We assessed statistical accuracy, ie, area under a receiver-operator characteristic curve (ROC); precision, ie, statistical goodness-of-fit; and actual impact on both risk-adjusted operative mortalities (RAOM) and performance rankings for 14 surgeons. The new model was slightly more accurate than the ready-made model (ROC, 0.78 versus 0.76; \( P < 0.05 \)), albeit not different from the recalibrated model (ROC, 0.77). The ready-made model showed poor fit between the predicted and observed results (\( P < 0.001 \)), leading to significant underestimation of RAOM (1.6 ± 0.2%) compared with the other strategies (2.5 ± 0.2%; \( P = 0.048 \)). Remodeling also changed the performance rankings among half the surgeons with higher RAOM.

Conclusions—Poorly calibrated risk algorithms can bias the calculation of RAOM and alter the results of surgeon-specific profiles. Any existing index used for risk assessment in cardiac surgery should be episodically recalibrated or compared with new models derived from local subjects to ensure that its performance remains optimal. (Circulation. 1999;99:2098-2104.)

Key Words: bypass ■ risk factors ■ mortality ■ prognosis

A long with the proliferation of public “report cards” on cardiac surgery in the United States, researchers have published many predictive rules or risk-adjustment algorithms for mortality, morbidity, and length of hospital stay after surgery. These risk indexes aim at identifying and weighting the patient characteristics that affect the probability of specific adverse outcomes. Indexes are then used retrospectively to adjust for case-mix differences among surgeons and centers when performance profiles are compiled. They are also used prospectively for patient counseling and for the identification of high-risk patient subgroups for special care or research.

See p 2067

Patient populations and the risk factors associated with adverse outcomes do change over time and also differ between centers. Thus, risk models derived and validated in 1 locale usually perform less well when applied in other settings or even to more contemporary patients in the same setting. Clinicians and managers considering the use of a risk index accordingly have 3 basic options:

1. They can use an existing, external index, knowing that the identified risk factors or at least the weights assigned to them may not be ideal for their patient populations (“ready-made”).
2. They can accept the risk factors in a published index but adjust the precision of the index to their own patient population by recalibrating the weights assigned to a published model.
3. They can derive an internal index from their own data (remodel).
In this article, we explore the implications of these options for assessment of operative mortality using a detailed data set with information on consecutive patients undergoing isolated CABG at 2 Toronto teaching hospitals. Specifically, we have compared the analytical strategies of recalibrating and remodeling against a ready-made rule developed by the authors for the Ontario Cardiac Care Network (CCN).

Methods

Data Sources

We examined clinical risk factors and operative mortality (OM) for all 7491 patients undergoing isolated CABG under the care of 14 surgeons at The Toronto Hospital (n = 5343) and Sunnybrook Health Science Center (n = 2148) between April 1, 1993, and December 31, 1996. Details of this database have been published previously.13 Data from earlier years were deliberately excluded because they were included in the multicenter data set used to derive and validate the CCN index. Twenty patients were missing ≥1 data elements used in the analyses.

Analysis

General Issues

Data were collected and managed in dBASEIV data sets. The SAS for PC18 and BMDP/DYN LR19 programs were used for statistical analysis. We were not seeking to recalibrate or rederive an index for external and general application. Thus, we forewent split-sample methods (ie, separate derivation and validation steps) with both the recalibrating and remodeling strategies described below.

Generating the Models

Ready-made: We used a 6-variable risk index developed for the Ontario CCN, drawing on all patients undergoing cardiac surgery in the province between April 1991 and March 1993. For isolated CABG outcomes, 1 variable (type of surgery) was set to its null value, leaving only 5 variables and their associated risk scores. The original regression coefficients for each variable were used to calculate patient-specific predicted probability (P) of OM from the formula:

\[ P_{OM} = \frac{e^{0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5}}{1 + e^{0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_5 X_5}} \]

where \( \beta_k \) is the constant and \( \beta_l \) is the regression coefficient for each level of a risk factor in the model that characterizes the patient.

Recalibrated: The 5 explanatory variables from the CCN index were included in logistic regression analyses of the 1993 to 1996 data set to reestimate mortality-specific regression coefficients and related risk scores. The predicted probability of OM as well as the total risk score for each patient was calculated as for the CCN model.1,20

Remodeled: The University of Toronto cardiac surgery registry covers a wide variety of potential risk factors.17 All explanatory variables with a univariate P value < 0.25, as well as those found commonly in other major risk indexes but failing to meet the critical alpha-level, were submitted to logistic regression analyses that used forward selection combined with backward elimination.21,22 The best logistic regression model was determined by 2 diagnostic criteria: the Hosmer-Lemeshow goodness-of-fit statistic21 and the area under the receiver-operator characteristic (ROC) curve.15,23,24 As in the CCN index, odds ratios were rounded to the nearest integer, and an additive risk index was created. Because of small numbers, risk scores > 18 were collapsed (ie, score ≥ 18 for OM).

Statistical Comparisons of Index Performance

Statistical accuracy, or model discrimination, was assessed with the area under the ROC curve23,24 for each model, with comparisons between models as described by Hanley and McNeil.24 Statistical precision, or model calibration, was evaluated by the Hosmer-Lemeshow goodness-of-fit statistic.21 We also plotted the mean predicted probability for OM against observed OM for each total risk score15,22 and performed a weighted linear regression to evaluate whether the relationship was overestimated or underestimated. A slope of 1 and intercept of 0 would indicate a perfect fit of predicted to observed outcomes.22 Differences in slopes and intercepts between the 3 regressions were evaluated by ANCOVA with pairwise comparisons as appropriate.

Clinically Salient Comparisons of Index Performance

Expected mortality for each surgeon for each model was calculated as the mean predicted probability of OM based on the prevalence of risk factors in his or her caseload. We calculated risk-adjusted OM (RAOM) by dividing the observed mortality by the expected mortality and then multiplying that ratio by the overall mortality rate in the study population (0.0226). This result can be interpreted as the mortality rate a surgeon would have if his or her case mix were similar to the average case mix in the study.25

Within each model, the difference between the mean observed mortality minus the mean expected mortality was evaluated by paired t test for the null hypothesis (H0) that the difference equaled zero.15 The differences in RAOM across 14 surgeons and 3 models were evaluated by ANOVA.

For each model, the surgeons were ranked from 1 (lowest RAOM) to 14 (highest RAOM) on the basis of how they ranked in the original CCN model. We examined qualitatively whether the ranking of surgeons changed and also calculated Spearman rank correlation coefficients (r_s) across models.

Results

Risk Factors and Risk Groups for the 3 Models

There were 169 deaths (2.26%). The prevalence of risk factors and their univariate association with OM are shown in Table 1.

The independent predictors of operative mortality for the new internal model were left ventricular grade, age group, previous bypass surgery, the timing of surgery, sex, triple-vessel disease, left main coronary artery disease, peripheral vascular disease, recent myocardial infarction, acute coronary insufficiency, and a history of hypertension. Given its recurrence as a risk factor in other published indexes, we forced preoperative renal insufficiency into the model, but its inclusion unfavorably affected both model discrimination and precision, possibly because of its low prevalence in our database or collinearity with other important predictors.

Table 2 contains the original ORs for OM and risk weights for the ready-made CCN model as well as the ORs, their 95% CIs, and risk weights from the recalibrated and remodeled indexes. When the CCN and remodeled indexes are compared, all 5 original risk factors do recur, but there are differences in weights, most notably for repeat operation and grade IV ventricular function.

Table 3 shows the number of patients defined by each risk score for each model and the observed OM for that model’s score. The addition of 4 explanatory variables to the remodeled index redefined “risk” in 87% of patients who bore at least 1 of the 4 conditions and resulted in a lower prevalence of both patients and OM at the lower risk scores.

Statistical Parameters of Index Performance

As shown in Table 4, the longer, remodeled index showed a small increment in accuracy over the original CCN index (ROC, 0.78 versus 0.76; P<0.05), with the recalibrated index...
between them (ROC, 0.77). The ready-made CCN model showed significantly poor fit between predicted and observed results \( (P < 0.001) \), whereas the other models had acceptable calibration.

Another method of assessing model fit is shown in Figure 1, which depicts the mean predicted probability of OM versus observed OM for each cumulative risk score. The original CCN model overestimated predicted OM (top panel). In contrast, for the recalibrated and new models, slopes appeared closer to unity and intercepts close to zero (data available on request). ANCOVA confirmed that there was a significant difference in intercepts among models. Pairwise comparisons showed that as with the ROC curve area, the significant difference arose only when the original CCN model and the newly remodeled index were compared \( (P < 0.0001) \).

OM: Observed, Expected, and Adjusted
We compared observed and expected mortality for each model. The observed minus expected mortality rate was significantly different from zero for the CCN model \((-1.04 \pm 0.4\%; P = 0.011) \) but not with recalibration \((0.17 \pm 0.3\%; P = 0.62) \) or remodeling \((0.21 \pm 0.3\%; P = 0.55) \). Intermodel comparisons by ANOVA confirmed that the CCN model had a significantly greater disparity between observed and predicted results compared with both the recalibrated \((P = 0.017 \text{ versus CCN}) \) and remodeled \((P = 0.014 \text{ versus CCN}) \) indexes.

As shown in Figure 2, this higher expected OM in the ready-made CCN model resulted in an underestimation of risk-adjusted OM \((P = 0.048 \text{ compared with the remodeled index}) \). RAOM was similar to unadjusted OM for the recalibrated and remodeled indexes.

Surgeon-Specific Rankings
Table 5 depicts the relative ranking of surgeons for each of the 3 models from lowest RAOM (rank = 1) to highest (rank = 14). Recalibration resulted only in surgeons 8 and 9 exchanging positions. Remodeling, however, resulted in surgeon 13 shifting up by 3 positions and surgeons 10 and 11 each moving down 2 ranks. Despite these latter shifts, the overall Spearman correlation coefficient showed a significant association between ranks for the CCN and new models \((r_s = 0.982, P = 0.012) \) because the positions of the first 7 surgeons were stable across models. However, examining the 7 higher-ranking surgeons revealed a diminished correlation \((r_s = 0.857, P = 0.07) \).

**Discussion**

We compared 3 possible strategies for assessing risk-adjusted outcomes of cardiac surgery: off-the-shelf use of a simple, multipurpose risk index; recalibrating that published index to ensure a better fit to the available data; and deriving a new, internal model with additional risk factors. The comparisons were effected in a clinical database of all isolated CABG patients undergoing operation between April 1, 1993, and December 31, 1996, at 2 large teaching hospitals in Toronto, Canada.

---

### TABLE 1. Univariate Results for the Prevalence and Operative Mortality of Risk Factors

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Prevalence, n (%)</th>
<th>OM, n (%)</th>
<th>( \chi^2 ) P Value</th>
<th>Unadjusted Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>5932 (79)</td>
<td>113 (1.9)</td>
<td>&lt;0.001</td>
<td>1.89</td>
</tr>
<tr>
<td>Female</td>
<td>1559 (21)</td>
<td>56 (3.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;65</td>
<td>4066 (54)</td>
<td>56 (1.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65–74</td>
<td>2653 (35)</td>
<td>73 (2.8)</td>
<td>1.99</td>
<td></td>
</tr>
<tr>
<td>( \geq 75 )</td>
<td>772 (10)</td>
<td>40 (5.2)</td>
<td>&lt;0.001</td>
<td>3.75</td>
</tr>
<tr>
<td>LV grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2428 (32)</td>
<td>22 (0.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3177 (42)</td>
<td>57 (1.8)</td>
<td>2.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1666 (22)</td>
<td>63 (3.8)</td>
<td>4.22</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>220 (3)</td>
<td>27 (12.3)</td>
<td>&lt;0.001</td>
<td>13.3</td>
</tr>
<tr>
<td>Timing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elective</td>
<td>4153 (55)</td>
<td>65 (1.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semiurgent</td>
<td>3035 (42)</td>
<td>83 (2.7)</td>
<td>1.69</td>
<td></td>
</tr>
<tr>
<td>Emergent</td>
<td>218 (3)</td>
<td>21 (9.6)</td>
<td>&lt;0.001</td>
<td>6.00</td>
</tr>
<tr>
<td>CABG redo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>7059 (94)</td>
<td>139 (2.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>432 (5.8)</td>
<td>30 (6.9)</td>
<td>&lt;0.001</td>
<td>3.45</td>
</tr>
<tr>
<td>Left main disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>6296 (84)</td>
<td>127 (2.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1195 (16)</td>
<td>42 (3.5)</td>
<td>0.001</td>
<td>1.75</td>
</tr>
<tr>
<td>Triple-vessel disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>2149 (29)</td>
<td>30 (1.4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5342 (71)</td>
<td>139 (2.6)</td>
<td>0.001</td>
<td>1.86</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>6398 (85)</td>
<td>121 (1.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1089 (15)</td>
<td>47 (4.3)</td>
<td>&lt;0.001</td>
<td>2.26</td>
</tr>
<tr>
<td>Hypertension</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>3733 (50)</td>
<td>63 (1.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3741 (50)</td>
<td>104 (2.8)</td>
<td>0.001</td>
<td>1.65</td>
</tr>
<tr>
<td>Recent MI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>5501 (74)</td>
<td>101 (1.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1977 (26)</td>
<td>67 (3.4)</td>
<td>&lt;0.001</td>
<td>1.89</td>
</tr>
<tr>
<td>Previous stroke/TIA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>6909 (92)</td>
<td>151 (2.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>579 (7.7)</td>
<td>18 (3.1)</td>
<td>0.151</td>
<td>1.41</td>
</tr>
<tr>
<td>Acute coronary insufficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>4884 (65)</td>
<td>89 (1.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>2606 (35)</td>
<td>80 (3.1)</td>
<td>0.001</td>
<td>1.72</td>
</tr>
<tr>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>5662 (76)</td>
<td>117 (2.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1824 (24)</td>
<td>51 (2.8)</td>
<td>0.067</td>
<td>1.33</td>
</tr>
<tr>
<td>Renal failure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>7227 (96)</td>
<td>149 (2.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>260 (3.5)</td>
<td>19 (7.3)</td>
<td>&lt;0.000</td>
<td>3.48</td>
</tr>
</tbody>
</table>
Our rationale was to offer some guidance to providers who must respond to the burgeoning literature on risk indexes. We accordingly discuss the implications of our findings in 3 areas.

Implications for Benchmarking Improvements Over Time
The original CCN model was derived and validated for earlier years. As such, it tended to overestimate the chances of postoperative death for high-risk patients, with the result that risk-adjusted outcomes improved. Experience with the Society of Thoracic Surgeons\textsuperscript{5,26} risk-adjustment algorithm has been similar. Their original 23-variable model was derived with Bayesian methods from data on tens of thousands of subjects and scores of centers. Nonetheless, in recent years, the model has predicted a rising probability of operative mortality owing to an increasing prevalence of high-risk patients, even as observed operative mortality has decreased.

If the goal of an outcomes analysis is to determine trends in mortality over time, then arguably a risk model derived and validated from an earlier period can be used, because it anchors practice historically and controls for the evolution of case mix. One limitation is that improved reportage of risk factors (“upcoding”) in contemporary groups of patients may lead to a spurious impression that risk-adjusted outcomes are improving. For example, critics have charged that this phenomenon, rather than the impact of public report cards, explains the improved outcomes of cardiac surgery in New York State.\textsuperscript{27}

Implications for Contemporaneous Quality Management or Patient Counseling
Setting aside temporal benchmarking, the usual goals of an outcomes analysis are contemporary quality management or risk prediction by, respectively, comparing the risk-adjusted outcomes of surgeons or centers or identifying patients in high-risk subgroups. For this purpose, our findings suggest that practitioners and managers should consider recalibrating an existing index or developing a new model with the data at hand.

Recall first that our ready-made CCN index was originally derived and validated in Ontario with data from 9 centers, including the 2 hospitals that contributed subsequent patients to the present study. Thus, it is perhaps not surprising that the CCN index still showed good performance in this study, with an ROC area of 0.76. However, the CCN index showed poor

### TABLE 2. Logistic Regression–Based Odds Ratios and Risk Weights for Operative Mortality

<table>
<thead>
<tr>
<th>Variable</th>
<th>CCN Model</th>
<th>Risk Weight*</th>
<th>Recalibrated OR (95% CI)</th>
<th>Risk Weight</th>
<th>Remodeled OR (95% CI)</th>
<th>Risk Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>LV grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.29</td>
<td>1</td>
<td>1.84 (1.1, 3.0)</td>
<td>2</td>
<td>1.81 (1.1, 3.0)</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1.95</td>
<td>2</td>
<td>3.75 (2.3, 6.2)</td>
<td>4</td>
<td>3.63 (2.2, 6.0)</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3.64</td>
<td>3</td>
<td>13.4 (7.3, 25)</td>
<td>13</td>
<td>11.6 (6.2, 22)</td>
<td>14</td>
</tr>
<tr>
<td>Age, y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65–74</td>
<td>1.88</td>
<td>2</td>
<td>1.82 (1.3, 2.6)</td>
<td>2</td>
<td>1.67 (1.2, 2.4)</td>
<td>2</td>
</tr>
<tr>
<td>≥75</td>
<td>2.59</td>
<td>3</td>
<td>3.36 (2.2, 5.2)</td>
<td>3</td>
<td>2.94 (1.9, 4.6)</td>
<td>3</td>
</tr>
<tr>
<td>CABG redo</td>
<td>3.22</td>
<td>2</td>
<td>3.16 (2.1, 4.8)</td>
<td>3</td>
<td>3.30 (2.1, 5.1)</td>
<td>3</td>
</tr>
<tr>
<td>Timing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semieurgent</td>
<td>1.45</td>
<td>1</td>
<td>1.28 (0.9, 1.8)</td>
<td>1</td>
<td>1.20 (0.8, 1.7)</td>
<td>1</td>
</tr>
<tr>
<td>Emergent</td>
<td>5.70</td>
<td>4</td>
<td>3.68 (2.1, 6.5)</td>
<td>4</td>
<td>3.45 (1.9, 6.2)</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>1.68</td>
<td>1</td>
<td>1.92 (1.4, 2.7)</td>
<td>2</td>
<td>1.82 (1.3, 2.6)</td>
<td>2</td>
</tr>
<tr>
<td>Triple-vessel disease</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>1.58 (1.0, 2.4)</td>
<td>2</td>
</tr>
<tr>
<td>Hypertension</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>1.45 (1.0, 2.0)</td>
<td>1</td>
</tr>
<tr>
<td>Left main disease</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>1.40 (1.0, 2.0)</td>
<td>1</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>1.39 (1.0, 2.0)</td>
<td>1</td>
</tr>
<tr>
<td>Constant</td>
<td>0.0100</td>
<td>0.00395</td>
<td>0.00221</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HL</td>
<td>0.45</td>
<td>0.354</td>
<td>0.599</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROC</td>
<td>0.75</td>
<td>0.755</td>
<td>0.778</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

HL indicates Hosmer-Lemeshow goodness-of-fit statistic; x, variable not submitted to the model.

The referent values are (in order) LV grade 1, age 65 years, primary CABG, elective timing, male sex, single- or double-vessel disease, no history of hypertension, no left main disease, no peripheral vascular disease.

*The risk weights for the “ready-made” CCN model were derived from an averaging of odds ratios across 3 separate outcomes: operative mortality, prolonged intensive care unit length of stay, and prolonged hospital length of stay. Odds ratios are presented here just for the operative mortality model.
calibration associated with overestimation of expected operative mortality, a feature of model performance that is undesirable for both prospective risk prediction and post hoc risk adjustment. Poor calibration presumably occurred not only because of temporal shifts in case mix but also because the CCN index was developed in a data set that combined ischemic and valvular heart disease, and we were applying it to an isolated CABG series. Parsonnet et al., for example, observed a deterioration in model performance when their predictive rule, developed in a data set that combined CABG with valve surgery, was subsequently tested for CABG and valve procedures separately. Similarly, the CCN index was derived to cover both OM and length-of-stay outcomes, and the recalibration here was outcome specific.

Recalibration was therefore a promising strategy in this context. More generally, it allows a group of practitioners to remain efficient in data collection, restricting their efforts to careful documentation of a limited number of prespecified variables. By reweighting these variables and fine-tuning the risk index, analysts may sometimes mitigate shifts in case mix and outcomes that occur either over time or as the index is applied to centers other than those from which it was derived.

Indeed, recalibration did lead to some improvements in model performance in this test case. Whereas the original CCN index demonstrated significantly poor fit with data from this new series of almost 7500 patients, the recalibrated index fit the data

### TABLE 3. Operative Mortality at Each Risk Score

<table>
<thead>
<tr>
<th>Risk Score</th>
<th>CCN Model, n (%)</th>
<th>Recalibrated, n (%)</th>
<th>Remodeled, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/703 (0.14)</td>
<td>1/703 (0.14)</td>
<td>0/150 (0.0)</td>
</tr>
<tr>
<td>1</td>
<td>10/1312 (0.76)</td>
<td>1/340 (0.29)</td>
<td>0/161 (0.0)</td>
</tr>
<tr>
<td>2</td>
<td>12/1533 (0.78)</td>
<td>12/1342 (0.89)</td>
<td>1/465 (0.22)</td>
</tr>
<tr>
<td>3</td>
<td>31/1379 (2.25)</td>
<td>5/897 (0.56)</td>
<td>2/622 (0.32)</td>
</tr>
<tr>
<td>4</td>
<td>23/1134 (2.03)</td>
<td>19/1195 (1.59)</td>
<td>6/848 (0.71)</td>
</tr>
<tr>
<td>5</td>
<td>28/804 (3.48)</td>
<td>22/996 (2.21)</td>
<td>6/926 (0.65)</td>
</tr>
<tr>
<td>6</td>
<td>32/383 (8.36)</td>
<td>19/618 (3.07)</td>
<td>13/922 (1.41)</td>
</tr>
<tr>
<td>7</td>
<td>13/140 (9.29)</td>
<td>21/587 (3.58)</td>
<td>17/916 (1.86)</td>
</tr>
<tr>
<td>8</td>
<td>6/50 (12.0)</td>
<td>13/303 (4.29)</td>
<td>20/746 (2.68)</td>
</tr>
<tr>
<td>9</td>
<td>8/29 (27.6)</td>
<td>13/132 (9.85)</td>
<td>16/536 (2.99)</td>
</tr>
<tr>
<td>10</td>
<td>5/17 (29.4)</td>
<td>6/100 (6.00)</td>
<td>15/418 (3.59)</td>
</tr>
<tr>
<td>11</td>
<td>...</td>
<td>2/29 (6.90)</td>
<td>16/268 (5.97)</td>
</tr>
<tr>
<td>12</td>
<td>...</td>
<td>2/12 (16.67)</td>
<td>12/137 (8.76)</td>
</tr>
<tr>
<td>13</td>
<td>...</td>
<td>6/67 (10.45)</td>
<td>7/92 (7.61)</td>
</tr>
<tr>
<td>14</td>
<td>...</td>
<td>4/36 (11.11)</td>
<td>5/52 (9.62)</td>
</tr>
<tr>
<td>15</td>
<td>...</td>
<td>8/71 (8.45)</td>
<td>4/51 (7.84)</td>
</tr>
<tr>
<td>16</td>
<td>...</td>
<td>15/56 (28.6)</td>
<td>3/39 (7.69)</td>
</tr>
<tr>
<td>17</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>≥18</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The addition of 4 variables to the remodeled index and the reweighting of key prognostic factors resulted in fewer patients at lower risk scores in the remodeled index and substantially lower observed OM compared with the other 2 models.

### TABLE 4. Evaluation of Accuracy and Precision of Predicted Versus Observed OM for Each of the Three Indices

<table>
<thead>
<tr>
<th></th>
<th>CCN Model</th>
<th>Recalibrated</th>
<th>Remodeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROC</td>
<td>0.760</td>
<td>0.770</td>
<td>0.780*</td>
</tr>
<tr>
<td>HL</td>
<td>&lt;0.001</td>
<td>0.246</td>
<td>0.383</td>
</tr>
</tbody>
</table>

HL indicates Hosmer-Lemeshow goodness-of-fit P value; a value of P < 0.05 indicates poor calibration.

*Different from the CCN model, P < 0.05.

Figure 1. Top, Results for mean predicted probability of OM as calculated by regression coefficients from original CCN model vs observed OM at each cumulative risk score (■). Parameter estimates for weighted linear regression analysis are given (intercept and slope), as well as R². Middle, Regression coefficients and risk weights were reestimated for those 5 variables in CCN model. Bottom, Results for more complex, 9-variable model. Dashed diagonal line represents a perfect fit between predicted and observed values.

Figure 2. Difference between observed (OBS) OM and risk-adjusted (RA) OM is depicted for each surgeon (solid lines). ▲, Mean observed OM. ■, Mean predicted (expected) outcome. ●. Mean risk-adjusted OM. Probability value refers to the difference between observed and risk-adjusted results. RAOM was different between the 3 models by ANOVA (P = 0.047).
TABLE 5. Risk-Adjusted Operative Mortality Ranking for Each Surgeon for Each Model

<table>
<thead>
<tr>
<th>Surgeon Rank</th>
<th>CCN Model</th>
<th>Recalibrated</th>
<th>Remodeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

The number for each surgeon was assigned according to his/her ranking position for RAOM rate as calculated by the original CCN model. The modeling strategy did not affect the relative ranking for the 7 surgeons with the lowest RAOM (Spearman correlation coefficient = 1.0, P < 0.0001). The type of model did, however, affect the ranking of surgeons with higher RAOM (Spearman correlation coefficient = 0.857, P = 0.069). Surgeon 13 moved up 3 places; surgeons 10 and 11 shifted down 2 places with remodeling.

well, and we avoided overestimating operative risk. The recalibrated model also showed discrimination similar to that of a new and more complex model and yielded similar relative ranks for most surgeons. However, for the higher RAOM surgeons, the new model did lead to some alterations in surgeon outcome rankings, an observation that underscores the potential practical importance of even small marginal improvements in model accuracy from a statistical standpoint.

In sum, for clinicians and managers who have developed their own index in the past or found an index that shows acceptable performance in their patient populations, episodic recalibration of that index may suffice. However, in those instances in which there are profound differences in case mix or event rates, it will be prudent to derive a new model with the data at hand.

How Many Risk Factors Are Enough?
Recently, the Society of Thoracic Surgeons published its updated risk model for 1995, developed from a database of >138,000 patients who underwent surgery at 374 hospitals throughout the United States and Canada. The model shows excellent accuracy but now requires 33 predictor variables. Apart from increased costs of data collection and increased risks of data “gaming” or random errors, the large numbers of explanatory variables also increase the chances of statistical overfitting and model instability when applied to specific centers.

In contrast, the original CCN model was designed to be parsimonious and robust for multicenter comparisons. The new model adds only 4 variables, bringing the total to 9 for isolated CABG. These factors are similar to those reported previously by our group and others and include those highlighted in recent guidelines from the Working Group Panel on the Cooperative CABG Database Project. Despite minor differences in surgeon rankings, this new model had performance characteristics similar to those of the recalibrated CCN model with only 5 variables. This latter result is consistent with our earlier findings on the limited marginal improvements in model performance with increasing numbers of predictor variables. Accurate and complete data collection on a constrained set of important variables appears to be a prudent strategy.

Conclusions
Our findings illustrate that temporal and intercenter differences in case mix make it difficult to achieve optimal predictive performance with ready-made risk indexes. This observation argues against the proliferation of published risk indexes in the clinical literature that either affirm well-known prognostic factors or add new variables with minimal marginal impact. We have also demonstrated that recalibration of existing indexes may sometimes be sufficient to ensure adequate risk prediction even when models are parsimonious. As a precaution, however, we suggest that centers collect data fastidiously on a modest-sized set of key variables such as those suggested by the Working Group Panel and undertake intermittent remodeling to ensure that emerging risk factors are not inadvertently overlooked.

Acknowledgments
J. Ivanov is supported by a fellowship from the Heart and Stroke Foundation of Canada. Dr. Tu is a Scholar and Dr. Naylor a Senior Scientist of the Medical Research Council of Canada. Drs. Tu and Naylor also receive personnel support from the Institute for Clinical Evaluative Sciences, which is funded by the Ontario Ministry of Health. The authors are deeply indebted to the cardiac surgeons at The Toronto Hospital and Sunnybrook Health Science Center who contributed patients to this study and allowed us access to their databases. The Toronto Hospital: Dr. Tirone E. David (chief), Dr. Ronald J. Baird, Dr. R.J. Cusimano, Dr. Christopher M. Feindel, Dr. Irvin H. Lipton, Dr. Lynda L. Mickelborough, Dr. Charles M. Peniston, Dr. Hugh E. Scully, and Dr. Richard D. Weiss. Sunnybrook Health Science Center: Dr. Bernard S. Goldman (chief), Dr. George T. Christakis, and Dr. Stephen E. Frenses. The authors also wish to acknowledge the superb contribution to data collection and management by Jeri Severs at the Sunnybrook Health Science Center and Susan Collins at The Toronto Hospital. We also wish to thank Dr. Donald Redeemeier, Dr. Antoni S.H. Basinski, Keyi Wu, and Kathy Sykora at the Clinical Epidemiology Unit of the Sunnybrook Health Science Center and the Institute for Clinical Evaluative Sciences of Ontario for statistical consultation.

References


Ready-Made, Recalibrated, or Remodeled?: Issues in the Use of Risk Indexes for Assessing Mortality After Coronary Artery Bypass Graft Surgery
Joan Ivanov, Jack V. Tu and C. David Naylor

Circulation. 1999;99:2098-2104
doi: 10.1161/01.CIR.99.16.2098

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1999 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/99/16/2098

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org/subscriptions/