CHAPTER TEN:
TESTING INTERNAL VALIDITY OF RISK ADJUSTMENT MODELS

For this study, the internal validity of a risk adjustment model is defined as how well it controls for differences in patient characteristics that would otherwise confound outcome comparisons across hospitals. A model that does not adequately control for such differences may generate biased and misleading estimates of risk-adjusted outcome rates. Internal validity of the risk-adjustment models was assessed in four basic ways: content validity, construct validity, discrimination, and calibration.

CONTENT VALIDITY

The models presented in Chapter Nine were reviewed with members of the AMI clinical advisory panel and outside consultants. The advisory panel included several cardiologists, one nurse researcher, and one coding professional with specialized expertise in the topic. They advised project staff about whether the models included appropriate covariates and whether the parameter estimates were consistent with previous research and experience in the field. Through this process, several variables with counterintuitive parameter estimates were eliminated from risk adjustment models. For example, hyperlipidemia was associated with a decreased risk of AMI mortality in an earlier risk model. This variable was eliminated from the final model because the negative parameter estimate was not consistent with previous research, and because selective underreporting of hyperlipidemia was strongly suspected. The clinical advisors and consultants generally agreed that the final models presented in Chapter Nine have content validity.

DISCRIMINATION

A model that distinguishes well between individuals who have poor outcomes and those who have good outcomes has excellent discrimination. A model with perfect discrimination would assign to every patient an expected probability of either zero or one; all persons with an expected probability of one, but no one with an expected probability of zero, would experience the outcome of interest. No model has perfect discrimination in the real world, but good models show substantial spread in the expected probability of the outcome (death) between those who actually experienced it and those who did not.

The most commonly used measure of discrimination is the c statistic, which represents the proportion of all randomly selected pairs of observations with different outcomes (e.g., one death and one survivor) in which the patient who died had a higher expected probability of death than the survivor.[1] The c statistic takes on values between 0 and 1.0; higher values indicate greater discrimination but there is no cutoff that distinguishes "adequate" from "inadequate" models. A value of 0.5 can be obtained by random selection.

Table 10.1 shows that the risk models for AMI mortality have c statistics of 0.774 for cases with no prior admissions and 0.759 for cases with one or more prior admissions.[2] These c statistics are based on Model A, which omitted demographic and clinical risk factors that may be unreliable or may reflect quality of care. As expected, Model B shows greater discrimination than Model A, with c statistics of 0.860 for cases with no prior admissions and 0.830 for cases with one or more prior admissions. This difference between the results for Model A and Model B is largely attributable to two powerful predictors that were used only in Model B: shock and pulmonary edema. These predictors were omitted from Model A because they may represent either in-hospital complications or associated conditions present on admission.

It is difficult to compare the performance of these risk models with that of models developed by other agencies evaluating hospital outcomes. Only a few such agencies report model performance measures. Pennsylvania's Health Care Cost Containment Council reported a c statistic of 0.772, using MedisGroups data elements in a specially designed model to predict coronary bypass mortality.[3] It has not reported c statistics for other subsets of patients. Using clinical data on coronary bypass patients from New York's Cardiac Surgery Reporting System, Hannan et al reported a c statistic of 0.787.[4] By comparison, the best he could achieve using administrative data

---

[1]Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29-36. The c statistic is equivalent to the area under a receiver operating characteristic curve, which represents a plot of sensitivity versus 1-specificity at various cutoff values for the predicted probability.

[2]These statistics are based on the complete 100% sample. A stricter test of model discrimination comes from applying a regression equation estimated using one sample to a set-aside, or validation sample. The resulting c statistics are virtually identical to those reported here for the "no-priors" models (0.774 for Model A; 0.860 for Model B), but slightly worse for the "priors" models (0.745 for Model A; 0.807 for Model B).

[3]The Pennsylvania Health Care Cost Containment Council. *Coronary Artery Bypass Surgery. A Technical Report*. Harrisburg, PA: November 1992.

[4]Hannan EL, Kilburn H, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. *JAMA* 1994; 271:761-766.

for the same patients was c = 0.74.[5] Cleveland Health Quality Choice has a very detailed data set with extensive clinical data; it reported c statistics of 0.85 to 0.92 from 5 risk-adjusted mortality models (including 0.89 for AMI cases).[6] Using Medicare claims from 84 randomly selected US hospitals to predict 30-day mortality, Krakauer et al reported a c statistic of 0.84.[7] This model was similar to that used by the Health Care Financing Administration to generate its reports on Medicare hospital mortality. No other agencies using administrative data to risk-adjust hospital outcomes have reported c statistics.

One recent study compared the ability of several severity indices to predict in-hospital mortality for AMI patients.[8] Among 775 patients treated either medically or surgically, the following c statistics were reported: 0.70 for APACHE II, 0.74 for Patient Management Categories (a system based on administrative data but designed to predict resource utilization), and 0.73 for MedisGroups. By using each index as an ordinal measure in a logistic regression model, these authors may have underestimated performance. Researchers at Queens University[9] used various commercial risk adjustment systems to predict 30-day and 60-day mortality among Medicare beneficiaries from six states. Across 23 DRG clusters, severity measures based on clinical data (MedisGroups, APACHE II, and Computerized Severity Index) had c statistics between 0.76 and 0.81. Severity measures based on administrative data (Acuity Index Method, DRGScale from Coded Staging, and Patient Management Categories) had c statistics between 0.72 and 0.75.

This summary demonstrates that the risk models developed as part of the California Hospital Outcomes Project compare favorably with others based on administrative data, but are probably inferior to those based on more detailed clinical data (e.g., APACHE III,[10] Cleveland Health Quality Choice).

---

[5]Hannan EL, Kilburn H Jr, Lindsey ML, Lewis R. Clinical versus administrative data bases for CABG surgery: Does it matter? *Medical Care* 1992; 30:892-907.

[6]Quality Information Management Corporation. Cleveland-Area Hospital Quality Outcome Measurements and Patient Satisfaction Report. Volume II. Cleveland, OH: Spring 1994.

[7]Krakauer H, Bailey RC, Skellan KJ, et al. Evaluation of the HCFA model for the analysis of mortality following hospitalization. *Health Services Research* 1992; 27:317-335.

[8]Alemi F, Rice J, Hankins R. Predicting in-hospital survival of myocardial infarction. *Medical Care* 1990; 28:762-775.

[9]Case Mix Research, Queens University. *Patient Classification Systems: An Evaluation of the State of the Art. Volume I.* Springfield, VA: National Technical Information Service, 1991.

[10]Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100:1619-1636.

CALIBRATION AND BIAS

Calibration is the extent to which observed outcome rates correspond to predicted rates across a set of defined strata. A well-calibrated model demonstrates excellent fit across a broad range of patient characteristics. Calibration may be a more relevant measure than discrimination when the purpose of a model is to predict outcome rates for groups of persons with similar characteristics (e.g., inpatients at the same hospital). By contrast, discrimination is more important if a model is being used to predict an individual's outcome and to make treatment decisions. The most commonly used measure of calibration is Hosmer and Lemeshow's chi square test,[11] which compares observed with predicted outcomes across several strata (e.g., 10) that are defined by increasing levels of risk.

Table 10.1 shows that the risk models for AMI mortality have marginally significant Hosmer-Lemeshow statistics, both among cases with no prior admissions ($\Pi 12$ = 19.01, p = 0.015) and among cases with one or more prior admissions (Π^2 = 14.92, p = 0.061). These chi-square statistics are based on Model A.[12] Model B shows a deterioration in this respect, as the Hosmer-Lemeshow statistics are 65.22 (p < 0.0001) among cases with no prior admissions and 27.24 (p < 0.0001) among cases with one or more prior admissions.[13] These models have relatively poor calibration because they overestimate the probability of death among the lowest-risk and highest-risk patients. Although attempts were made to correct this problem by testing additional interaction terms, this effort had limited success. In developing Model B, the focus was on maximizing discrimination at the expense of other model characteristics.

These tests confirm that most AMI risk models developed as part of the California Hospital Outcomes Project meet generally accepted standards of calibration. With the exception of AMI Model B, these models do not demonstrate a significant, consistent pattern of bias across risk strata.

Bias tests also were performed for a variety of other patient characteristics that were deliberately omitted from the risk-adjustment models or specified in

---

[11]Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: John Wiley & Sons, 1989.

[12]These statistics are based on the complete 100% sample. A better test of model calibration comes from applying a regression equation estimated using 60% of the cases to the remaining 40% validation sample. This procedure generated nonsignificant Hosmer-Lemeshow statistics, both for cases with no prior admissions (Π^2 = 6.58, p = 0.58) and for cases with one or more prior admissions (Π^2 = 10.01, p = 0.26).

[13]These statistics are based on the complete 100% sample, but similar results were obtained from the set-aside sample. This procedure generated Hosmer-Lemeshow statistics of 32.13 for cases with no prior admissions (p < 0.0001) and 21.94 for the remaining cases (p = 0.005).

a particular manner. None of these models show bias related to age, race, or date of admission. AMI Model A shows bias related to the source and type of admission, due to the deliberate omission of these variables from Model A.[14]

Bias testing therefore confirmed that the risk-adjustment models developed for the California Hospital Outcomes Project are relatively free from bias due to temporal and demographic factors. Of course, substantial bias due to unmeasured clinical factors is likely.

[14]Model B eliminates these biases, but may improperly underestimate true differences in risk-adjusted outcomes across hospitals by adjusting for source and type of admission.

Table 10.1: Goodness-of-fit tests for AMI mortality models				
	Priors		*No Priors*	
	Model A	*Model B*	*Model A*	*Model B*
Number of cases	5,442	5,415	62,570	62,220
Number of deaths	1,044	1,039	7,803	7,763
Death rate, %	19.18	19.19	12.47	12.48
Model chi square	721.73	1,276.49	6,775.57	13,630.56
df	13	25	24	44
p value	0.0001	0.0001	0.0001	0.0001
C statistic	0.759	0.830	0.774	0.860
Hosmer Lemeshow statistic	14.92	27.24	19.01	65.22
df	8	8	8	8
p value	0.0607	0.0006	0.0148	0.0001