Risk-Adjusting Acute Myocardial Infarction Mortality: Are APR-DRGs the Right Tool?

Patrick S. Romano and Benjamin K. Chan

Objective. To determine if a widely used proprietary risk-adjustment system, APR-DRGs, misadjusts for severity of illness and misclassifies provider performance.

Data Sources. (1) Discharge abstracts for 116,174 noninstitutionalized adults with acute myocardial infarction (AMI) admitted to nonfederal California hospitals in 1991–1993; (2) inpatient medical records for a stratified probability sample of 974 patients with AMIs admitted to 30 California hospitals between July 31, 1990 and May 31, 1991.

Study Design. Using the 1991–1993 data set, we evaluated the predictive performance of APR-DRGs Version 12. Using the 1990/1991 validation sample, we assessed the effect of assigning APR-DRGs based on different sources of ICD-9-CM data.

Data Collection/Extraction Methods. Trained, blinded coders reabstracted all ICD-9-CM diagnoses and procedures, and established the timing of each diagnosis. APR-DRG Risk of Mortality and Severity of Illness classes were assigned based on (1) all hospital-reported diagnoses, (2) all reabstracted diagnoses, and (3) reabstracted diagnoses present at admission. The outcome variables were 30-day mortality in the 1991–1993 data set and 30-day inpatient mortality in the 1990/1991 validation sample.

Principal Findings. The APR-DRG Risk of Mortality class was a strong predictor of death ($c = .831$–$.847$), but was further enhanced by adding age and sex. Reabstracting diagnoses improved the apparent performance of APR-DRGs ($c = .93$ versus $c = .87$), while using only the diagnoses present at admission decreased apparent performance ($c = .74$). Reabstracting diagnoses had less effect on hospitals' expected mortality rates ($r = .83$–$.85$) than using diagnoses present at admission instead of all reabstracted diagnoses ($r = .72$–$.77$). There was fair agreement in classifying hospital performance based on these three sets of diagnostic data ($\kappa = 0.35$–0.38).

Conclusions. The APR-DRG Risk of Mortality system is a powerful risk-adjustment tool, largely because it includes all relevant diagnoses, regardless of timing. Although some late diagnoses may not be preventable, APR-DRGs appear suitable only if one assumes that none is preventable.

Key Words. Risk adjustment, APR-DRG, risk of mortality, severity of illness, acute myocardial infarction, report card, hospital performance

In the increasingly competitive marketplace for hospital services, employers, health plans, and government agencies have started generating comparative data on hospital outcomes and costs (Pennsylvania Health Care Cost Containment Council 1996; Office of Statewide Health Planning and Development 1996; New York State Department of Health 1996). These data must be risk-adjusted to create a level playing field on which quality differences can be separated from differences attributable to severity of illness. Numerous risk-adjustment systems are promoted for this purpose, but none is clearly superior and different systems may give quite different answers (Iezzoni 1997a).

ALL PATIENT REFINED–DIAGNOSIS-RELATED GROUPS (APR-DRGS)

All Patient Refined–Diagnosis-Related Groups have become one of the most popular commercial systems for severity adjustment using hospital discharge data. The system's vendor, 3M Health Information Systems, claims that APR-DRGs are "the most comprehensive, clinically accurate severity of illness and risk of mortality product available." Such marketing efforts have led several states to use APR-DRGs to compare hospital performance. After abandoning the more costly MedisGroups system, for example, agencies in both Iowa and Colorado selected APR-DRGs to risk-adjust charge and mortality data from acute care hospitals (Iezzoni 1997b). APR-DRGs are also being used in Florida, Utah, and Michigan to risk-adjust mortality, length of stay, or inpatient charges for public report cards.[1]

APR-DRG software assigns three descriptors to each case: (1) the "base APR-DRG," which for adults generally represents a combination of adjacent Medicare DRGs split by age, death, comorbidities or complications; (2) the Severity of Illness (SOI) class; and (3) the Risk of Mortality (ROM) class. Severity of illness is defined as the extent of organ system derangement or

---

This work was performed under Inter-Agency Agreement 95-6225 with the Office of Statewide Health Planning and Development, State of California.

Address correspondence to Patrick S. Romano, M.D., M.P.H., University of California, Davis, School of Medicine Division of General Medicine, PSSB - Suite 2400, 4150 V Street, Sacramento, CA 95817. Dr. Romano is Associate Professor of Medicine and Pediatrics and Benjamin K. Chan, M.S. is a Statistician, Division of General Medicine and Center for Health Services Research in Primary Care, UC Davis School of Medicine. This article, submitted to *Health Services Research* on March 31, 1998, was revised and accepted for publication on March 29, 1999.

physiologic decompensation, whereas risk of mortality is the likelihood of in-hospital death (Averill, Muldoon, Vertrees, et al. 1997). The SOI and ROM classes (minor, moderate, major, and extreme) are determined separately based on secondary diagnoses and interactions between these diagnoses and age, principal diagnosis, and selected procedures. Users are instructed to use the SOI class for "evaluating resource use or establishing patient care guidelines" and the ROM class for "evaluating patient mortality." Version 12.0 included 384 base APR-DRGs and 1,530 severity-stratified APR-DRGs (3M Health Information Systems 1995).

In recent studies, APR-DRG SOI classes discriminated better in predicting inpatient mortality after acute myocardial infarction (AMI)(Iezzoni, Ash, Shwartz, et al. 1996) and coronary artery bypass (CABG) surgery (Iezzoni, Ash, Shwartz, et al. 1998) than did systems that require detailed clinical abstraction. APR-DRG SOI classes performed better than any other method based on discharge abstracts for patients with stroke (Iezzoni, Shwartz, Ash, et al. 1995) and CABG (Iezzoni, Ash, Shwartz, et al. 1998), and about as well as two competing products for patients with AMI (Iezzoni, Ash, Shwartz, et al. 1996) and pneumonia (Iezzoni, Shwartz, Ash, et al 1996). APR-DRGs did not include ROM classes when these studies were performed; this new measure might be even more powerful because it was designed specifically to predict in-hospital mortality.

However, predictive power is only one criterion for evaluating risk-adjustment methods. The ideal tool for comparative performance evaluation would adjust for conditions that reflect the patient's severity of illness at admission or the natural history of the patient's illness, but would not adjust for complications that could have been averted or ameliorated with optimal medical care. The *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) does not distinguish comorbidities from complications; researchers have made creative attempts to do so (Iezzoni, Daley, Heeren, et al. 1994; Brailer et al. 1996), but the validity of these efforts is unclear. As a result, risk-adjustment tools based on ICD-9-CM, such as APR-DRGs, may misadjust for severity of illness and implicitly excuse hospitals for iatrogenic complications.

The California Hospital Outcomes Project's Acute Myocardial Infarction Validation Study (Office of Statewide Health Planning and Development [OSHPD] 1996) offered an opportunity to test this hypothesis. Hospital discharge data submitted to California's OSHPD were reabstracted, with attention to the timing of each diagnosis, so that comorbidities and complications could be distinguished. By applying APR-DRG software to the resulting lists

of diagnoses, we explored whether adding reabstracted diagnoses and removing complications from the risk-adjustment process would affect classification of hospital performance.

METHODS

Data

Two data sets, both derived originally from OSHPD's Patient Discharge Data Set, were used in this study: (1) the California Hospital Outcomes Project 1991–1993 AMI subset and (2) the 1990/1991 AMI Validation Study data set. The Patient Discharge Data Set includes all discharges from licensed nonfederal hospitals in California. The variables collected include a hospital facility number; the patient's date of birth, social security number (SSN), zip code, race, sex, and disposition; the dates of admission and discharge; the principal source of payment and total charges; the source and type of admission; the principal diagnosis and up to 24 other diagnoses; the principal procedure and up to 20 other procedures; and up to four codes for external causes of injury.

The inclusion and exclusion criteria for both data sets are described in detail elsewhere (Office of Statewide Health Planning and Development 1996; Romano et al. 1997).[2] In brief, we selected acute care discharges with a principal diagnosis of 410.x0 or 410.x1, or certain related principal diagnoses (427.1, 427.4x–427.5, 429.5–429.7x, 429.81, 518.4, 780.2, or 785.51) with a secondary diagnosis of 410.x0 or 410.x1. We excluded patients less than 18 years of age at admission and patients admitted from skilled nursing or intermediate care facilities. When patients were transferred between hospitals, we linked sequential records and ascribed the patient's 30-day outcome to the originating hospital.[3] After linkage, we excluded patients who were discharged alive less than two days after admission (three days in the 1990/1991 AMI Validation data set), unless they went to another hospital or left against medical advice. We also excluded patients who had a prior AMI (any diagnosis of 410.x0 or 410.x1) within eight weeks of the index admission, patients involved in a transport accident (E800.x–E848), and in the AMI Validation data set only, patients with a secondary diagnosis of congenital or rheumatic aortic stenosis (746.3, 395.0, 395.2, 396.0, 396.2, or 396.8).[4]

Finally, we identified hospitals with evidence of substandard ICD-9-CM coding, based on the distribution of patient characteristics across hospitals. The specific criteria applied to the 1991–1993 AMI subset were (1) prevalence

of subendocardial infarction (410.7x) less than 12 percent, (2) prevalence of other/unspecified site (410.8x–410.9x without 410.0x–410.7x) more than 28 percent, (3) prevalence of hypertension (401.x–405.xx) less than 14 percent, and (4) prevalence of congestive heart failure (425.x, 428.x) less than 17 percent. Hospitals were excluded if the exact probability of any of these findings was less than approximately .00003.[5] The combined effect of these criteria was to exclude 27 hospitals with 2,127 AMI patients in 1991, 17 hospitals with 1,068 AMI patients in 1992, and 13 hospitals with 494 AMI patients in 1993. We applied similar criteria to the 1990/1991 Validation data set, except that we also excluded 16 hospitals that transferred at least 20 percent of their AMI discharges to nonreporting (e.g., out-of-state or federal) facilities.[6] All of our criteria were validated through a confidential survey of the excluded hospitals.

The California Hospital Outcomes Project 1991–1993 AMI subset included all qualifying AMI cases admitted between January 1, 1991 and December 1, 1993. The final data set included 116,174 unique records, which were linked with death certificates by OSHPD staff.[7] We used this data set for overall analyses of APR-DRG performance among AMI patients.

The AMI Validation Study was designed to evaluate the validity of using discharge abstracts to estimate risk-adjusted AMI mortality rates for California hospitals. Our sampling frame was the 1990/1991 sample described earlier, which included 31,512 qualifying AMI admissions between July 31, 1990 and May 31, 1991. In the first stage of probability sampling, we stratified hospitals by their AMI volume (<50, 50–117, or >117) and mortality rating in the 1993 report (better than expected, worse than expected, or neither better nor worse). We excluded low-volume hospitals to improve efficiency. We randomly selected six medium-volume and four high-volume hospitals from each mortality stratum; one hospital refused to participate and was replaced by a randomly selected alternate. In the second stage, we stratified AMI patients at each of the 30 sampled hospitals by outcome (in-hospital death within 30 days of admission). We oversampled deaths to produce a target death rate of 26 percent, or twice the statewide rate, within each hospital stratum. About 133 deaths and 47 survivors were randomly sampled from each stratum, generating a desired sample of 1,065 cases.

We requested a complete photocopy of the record for each case, and received 1,005 (94.4 percent). After careful review, we excluded 31 records with no evidence of AMI. Experienced accredited records technicians/certified coding specialists, blinded to the original discharge abstracts, used ICD-9-CM to reabstract all diagnosis and procedure codes from the remaining

974 records. A nurse or physician reviewer then verified the ICD-9-CM diagnoses and abstracted clinical data elements from each record. All coders and clinician reviewers received detailed written guidelines and intensive training. They entered their findings directly into a computerized data entry system equipped with built-in error and logic checks. Supervisors monitored productivity and reviewed at least 5 percent of each abstractor's records. We used this data set to explore the sensitivity of APR-DRGs to the timing of diagnoses and the quality of diagnostic coding.

Outcomes

The primary outcome variable in the California Hospital Outcomes Project 1991–1993 AMI subset was 30-day mortality, regardless of site. We chose this outcome variable because it is unbiased by length of stay and therefore has the greatest inherent validity for comparative performance studies. We obtained outcome information from either the discharge abstract or a linked death certificate. Overall, 13,976 (11.7 percent) patients were reported as 30-day deaths in both data sets; 936 (0.8 percent) were reported as inpatient, 30-day deaths on hospital discharge abstracts but were not confirmed by vital statistics; and 2,501 (2.1 percent) had linked death certificates within 30 days but were discharged alive from the hospital.

The primary outcome variable in the 1990/1991 AMI Validation Study data set was inpatient 30-day mortality, regardless of hospital. We chose this outcome variable because it is unbiased by hospital transfer and referral practices; but it does not require linkage of hospital discharge abstracts with death certificates (which became available after the Validation Study was completed). We obtained outcome information from either the index discharge abstract or subsequent linked abstracts.

Models

We generated a series of logistic regression models to estimate the probability of death for each case. Each model was estimated three times, based on different sets of ICD-9-CM diagnoses: (1) the codes originally reported to OSHPD, (2) the codes we reabstracted, and (3) the reabstracted codes for diagnoses that were present at admission. We based this determination on whether the diagnosis was documented or suggested in any prehospital, emergency room, or admission note. In a separate analysis, we tested the effect of labeling all diagnoses documented or suggested on the day of arrival or the following day as "present at admission." These two analyses produced identical results.

Following Iezzoni and colleagues (1995, 1996, 1998), we constructed models that included (1) the APR-DRG ROM class alone; (2) the ROM class with dummy variables representing a cross-classification of patients by sex and eight age categories (18–44, 45–54, 55–64, 65–69, 70–74, 75–79, 80–84, and 85 years of age or older); and (3) the ROM class with variables representing female gender, age (in years), the interaction between age and female gender, and linear age less than 35 years (piecewise regression). We present models (1) and (3) because they maximize discrimination and simultaneously conserve degrees of freedom; the results of models (2) and (3) differ by .001 or less.

We reestimated the same set of models using the APR-DRG SOI class, and then the cross-classification of both SOI and ROM classes, in place of the ROM class alone. These analyses provide direct comparability with previous literature (Iezzoni et al. 1995, 1996, 1998) and test whether the SOI class provides additional information for predicting mortality relative to the ROM class alone.

We tested three different specifications of the ROM and SOI classes in these models. First, we entered each ROM class, SOI class, or combination thereof (e.g., ROM = 1 and SOI = 1) as a dummy variable. This method optimized predictive power because it fit each model to the particular characteristics of its cases. However, we could apply it only to the subset of cases in APR-DRG 121 (medically managed AMI), because other APR-DRGs (e.g., CABG, pacemaker insertion) were too infrequent to estimate mortality reliably using our AMI data.

Second, we used the entire 1992–1993 California Patient Discharge Data Set as a standard to estimate the probability of death for patients with each combination of APR-DRG values. We entered the logit of this probability as a single independent variable; its regression coefficient in a perfectly calibrated model would equal one. We prefer this approach because it permits the inclusion of relatively infrequent APR-DRGs and simulates real-world use of the product (with comparative norms generated by HCIA Inc.). Because the HCIA benchmarks were not available under our licensing agreement, we generated our own benchmarks using OSHPD's entire data set.

Third, we used the same logit probabilities in a "reduced" data set limited to cases in APR-DRG 121. Using this set of models, we separated the effect of studying only medically managed cases from the effect of using logit probabilities instead of dummy variables to estimate the impact of APR-DRGs.

Statistical Methods

We used unweighted logistic regression in all analyses involving the California Hospital Outcomes Project 1991–1993 AMI subset. To account for the complex stratified sampling scheme, we used weighted logistic regression in all analyses involving the 1990/1991 AMI Validation data set. Each case was weighted by the inverse of its sampling probability, after adjustment for nonresponse, so that the weighted sample would resemble the population from which it was drawn.

We evaluated the predictive performance of these models using the *c*-statistic, which represents the proportion of all randomly selected pairs of observations with different outcomes in which the patient who died had a higher expected probability of death than the survivor (Hanley and McNeil 1982). The *c*-statistic is equivalent to the area under a receiver-operating characteristic curve. We also used the R^2-statistic, which represents one minus the ratio of model deviance to the deviance of an intercept-only model (Hosmer and Lemeshow 1989). This formulation is better than the ratio of squared residuals for evaluating logistic models, because it always increases as more independent variables are added and equals one for the saturated model (Cameron and Windmeijer 1997).

If the purpose of a model is to predict outcome rates for groups of individuals (e.g., inpatients at the same hospital), calibration may be an even more relevant measure of validity. Calibration is the extent to which observed outcome rates correspond to predicted outcome rates across a set of defined strata. In analyses based on the California Hospital Outcomes Project 1991–1993 AMI subset, we evaluated calibration using Hosmer and Lemeshow's test (1989), which compares observed and predicted outcomes across ten strata defined by increasing levels of risk. Although this test may be sensitive to the specific algorithm used to define interdecile cutpoints, no other goodness-of-fit test proved clearly superior in simulations (Hosmer et al. 1997). In analyses based on the 1990/1991 AMI Validation Study data set, we compared observed and predicted outcomes without estimating a test statistic.

We assessed the effect of different risk-adjustment methods on hospital classification by aggregating probabilities of death at the hospital level and dividing each hospital's number of observed deaths by its number of expected deaths to obtain an indirectly standardized mortality ratio (ISMR). We estimated the population variance of this parameter (Levy 1991) and flagged hospitals with ISMR values significantly greater or less than one, based on the assumption that our weighted sample was representative of their population of AMI patients.

We then estimated the intercorrelations among patient-level probabilities of death, and hospital-level expected mortality rates, from different risk-adjustment models. Both Spearman rank order and Pearson correlation coefficients were estimated; only the Pearson coefficients are reported here for comparison with Iezzoni, Ash, Shwartz, et al. (1996). In patient-level analyses involving the 1990/1991 AMI Validation data set, each case was weighted by the inverse of its sampling probability (after adjustment for nonresponse). We used the kappa statistic to evaluate the extent of agreement, corrected for chance, in classifying hospital performance as better than expected, worse than expected, or neither (Soeken and Prescott 1986).

RESULTS

Sample Characteristics

The California Hospital Outcomes Project 1991–1993 AMI subset included 116,174 cases. Forty-eight of these cases had to be excluded because there were zero deaths with the same APR-DRG values in OSHPD's benchmark data set. As a result, the logit probability of death for these 48 cases was undefined.

The 1990/1991 AMI Validation Study data set included 974 cases diagnosed as having an acute myocardial infarction. Three of these cases in APR-DRG 115 (permanent cardiac pacemaker implantation) were excluded because the logit probability of death from OSHPD's benchmark data set was undefined.

Overall APR-DRG Performance

Overall APR-DRG performance was evaluated using the California Hospital Outcomes Project 1991–1993 AMI subset. Table 1 shows the *c*-, R^2-, and Hosmer-Lemeshow χ^2-statistics from models including all eligible cases and cases in APR-DRG 121 (medically managed AMI) only. For cases in APR-DRG 121, we report separately the results of the logit probability method (where a single variable represents the entire class effect) and the dummy method (where a separate dummy variable represents each class). The latter two methods generated virtually identical *c*- and R^2-statistics, but the Hosmer-Lemeshow χ^2-statistic was generally larger when the logit probability was used instead of dummy variables to represent the effect of APR-DRG classes. This statistic was also larger when all AMI cases were used instead of just medically managed cases.

Table 1: Performance Characteristics of Risk-Adjustment Models with APR-DRGs, Using the California Hospital Outcomes Project 1991–1993 AMI Subset (N =116,126)

Model Components	*Performance Measure*	*All AMIs Logit Probability*	*APR-DRG 121 Logit Probability*	*APR-DRG 121 Dummy Variables*
Risk of Mortality	c (95% CI)	.847 (.844–.851)	.831 (.827–.834)	.831 (.827–.834)
	R^2	.288	.277	.277
	H-Lχ^2 (8 df)	38.03**	16.01*	13.14
Severity of Illness	c (95% CI)	.818 (.814–.821)	.803 (.799–.806)	.803 (.799–.806)
	R^2	.230	.210	.210
	H-Lχ^2 (8 df)	24.39**	13.13	11.15
Risk of Mortality + Severity of Illness	c (95% CI)	.854 (.851–.857)	.841 (.837–.845)	.841 (.838–.845)
	R^2	.294	.284	.284
	H-Lχ^2 (8 df)	41.86**	11.98	4.96
Risk of Mortality + Age/Sex	c (95% CI)	.870 (.867–.873)	.859 (.856–.863)	.860 (.857–.864)
	R^2	.319	.309	.311
	H-Lχ^2 (8 df)	90.30**	88.41**	21.71**
Severity of Illness + Age/Sex	c (95% CI)	.844 (.841–.847)	.834 (.831–.838)	.835 (.832–.838)
	R^2	.249	.238	.238
	H-Lχ^2 (8 df)	139.26**	102.96**	166.24**
Risk of Mortality + Severity of Illness + Age/Sex	c (95% CI)	.874 (.871–.876)	.864 (.861–.867)	.865 (.862–.869)
	R^2	.324	.315	.318
	H-Lχ^2 (8 df)	95.94**	83.18**	22.00**

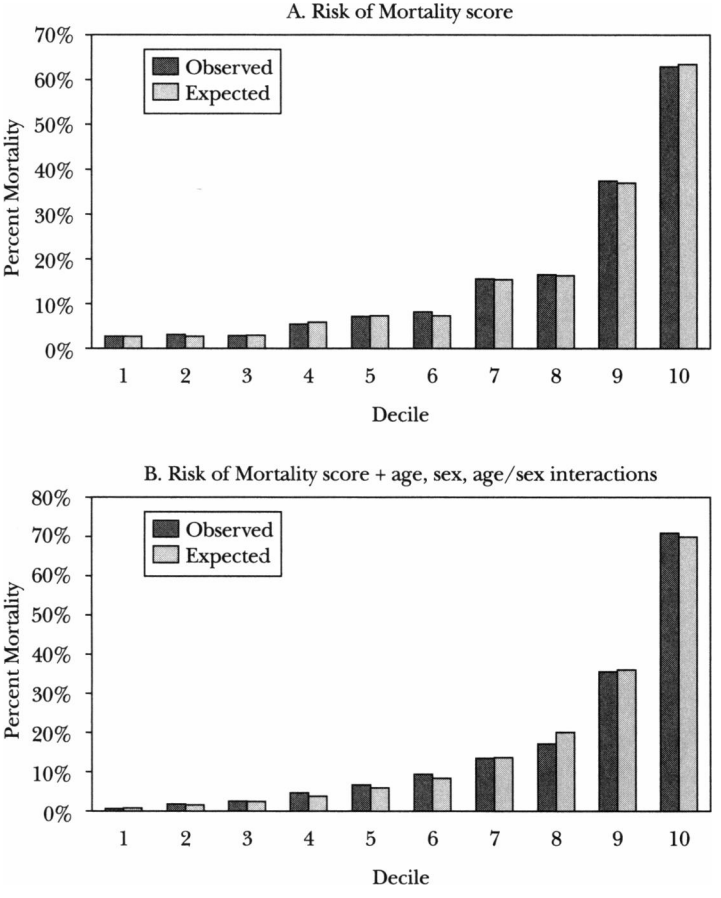
*Significant at 5% level; **significant at 1% level.

The Risk of Mortality class was an excellent predictor of death among AMI patients (c = .831–.847). It discriminated better (i.e., higher c- and R^2-statistics), with comparable calibration (Hosmer-Lemeshow χ^2-statistic), than the severity of illness class that Iezzoni and colleagues had previously evaluated. Use of the ROM and SOI classes together, as shown in Table 1, was slightly better than using the ROM class alone, but was much better than using the SOI class alone. Adding age, sex, and the interaction between them substantially improved the discrimination of all APR-DRG models. However, the calibration of these models was consistently worse than the calibration of models without demographic information (Figure 1).

Effect of Using Reabstracted ICD-9-CM Codes

Assigning APR-DRGs based on the originally reported diagnoses in our 1990/1991 AMI Validation data set, we generated similar estimates of predictive performance (Table 2), although the superiority of ROM classes over

Figure 1: Observed Versus Expected Mortality by Decile of Risk, Using the APR-DRG Risk of Mortality Score (with or without Age, Sex, and Age/Sex Interactions) in the California Hospital Outcomes Project 1991–1993 AMI Subset, APR-DRG 121 Cases ($N = 92{,}159$)



SOI classes was even more apparent ($c = .87$ versus $c = .80$ among all AMI cases). The difference in model performance between the 1990/1991 AMI Validation data set and the California Hospital Outcomes Project 1991–1993 AMI subset represents the combined effects of sampling variation and variation over time.

Blinded reabstraction of medical records by specially trained ICD-9-CM coders led to an increase in the predictive power of APR-DRGs, using

Table 2: Performance Characteristics of Risk-Adjustment Models with APR-DRGs, Using the 1990/1991 AMI Validation Study Data Set ($N = 971$ cases)

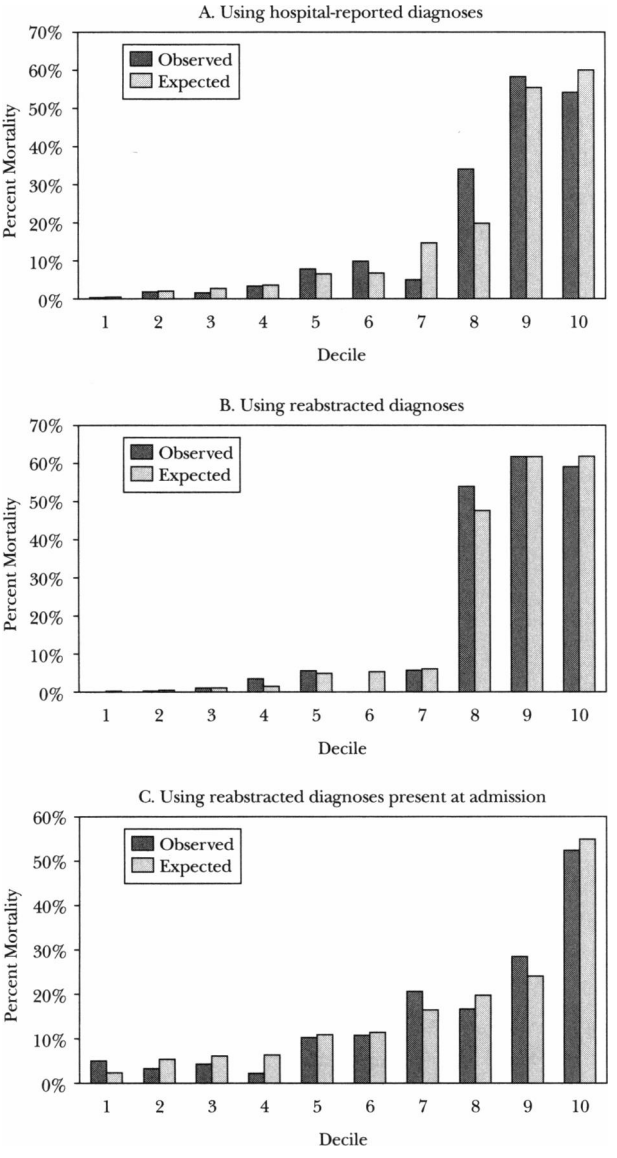
Model Components	*Performance Measure*	*All AMIs Logit Probability*	*APR-DRG 121 Logit Probability*	*APR-DRG 121 Dummy Variables*
Risk of Mortality Original ICD-9-CM codes	c (95% CI) R^2	.87 (.84–.90) .31	.86 (.82–.89) .28	.86 (.82–.89) .28
Severity of Illness Original ICD-9-CM codes	c (95% CI) R^2	.80 (.77–.83) .20	.80 (.77–.84) .17	.80 (.77–.84) .19
Risk of Mortality Reabstracted ICD-9-CM codes	c (95% CI) R^2	.93 (.91–.95) .45	.92 (.90–.94) .44	.92 (.90–.94) .45
Severity of Illness Reabstracted ICD-9-CM codes	c (95% CI) R^2	.83 (.81–.86) .26	.83 (.80–.86) .24	.83 (.80–.86) N/A*
Risk of Mortality Reabstracted ICD-9-CM codes present at admission	c (95% CI) R^2	.74 (.70–.78) .15	.75 (.71–.79) .12	.75 (.71–.79) .13
Severity of Illness Reabstracted ICD-9-CM codes present at admission	c (95% CI) R^2	.76 (.73–.80) .15	.79 (.75–.82) .13	.79 (.75–.82) .13

*Deviance could not be calculated because the model did not converge using maximum likelihood estimation.

either ROM classes ($c = .93$ versus $c = .87$ among all AMI cases) or SOI classes ($c = .83$ versus $c = .80$ among all AMI cases). These differences reflect the additional ICD-9-CM codes that hospitals should have reported to OSHPD, but did not. The use of reabstracted diagnoses did not have a consistent effect on calibration (Figure 2). These findings were consistent whether we included all AMI cases or cases that were assigned only to APR-DRG 121.

Using diagnoses reabstracted from only prehospital, emergency room, and admission notes ("present at admission") to assign APR-DRGs led to a substantial drop in the predictive performance of the ROM class ($c = .74$ versus $c = .93$ among all AMI cases) and a smaller drop in the predictive performance of the SOI class ($c = .76$ versus $c = .83$ among all AMI cases). Using only diagnoses present at admission, the SOI class predicted mortality as well as the ROM class did, even though the latter variable was specifically designed for this purpose. The use of diagnoses present at admission did not have a consistent effect on calibration (Figure 2). These findings were consistent whether we included all AMI cases or cases only in APR-DRG 121.

Figure 2: Observed Versus Expected Mortality by Decile of Risk, Using the APR-DRG Risk of Mortality Score on Hospital-Reported Diagnoses, Reabstracted Diagnoses, and Reabstracted Diagnoses Present at Admission in the 1990/1991 AMI Validation Study Data Set, APR-DRG 121 Cases ($N = 922$)



Effect on Probabilities of Death

In the California Hospital Outcomes Project 1991–1993 AMI subset, the probabilities of death estimated using the APR-DRG ROM class alone were highly correlated ($r > .93$) with those estimated using the ROM class plus the SOI class, age and sex, or both sets of extra variables; and they were moderately correlated ($r = .77$) with those estimated using the SOI class alone. The correlations between paired estimates of the probability of death were considerably weaker when we used different sets of ICD-9-CM diagnoses to assign APR-DRGs in the 1990/1991 AMI Validation data set. The weighted Pearson correlations between probabilities estimated using all reabstracted diagnoses and those estimated using only diagnoses present at admission were .70 or lower for the ROM class but .80 or higher for the SOI class (Table 3).

The same trends were noted when we examined the correlations between paired severity models at the hospital level (Table 4). The use of

Table 3: Weighted Pearson Correlations Between Patient-Level Probabilities of Death from Risk-Adjustment Models with APR-DRGs, Using All Cases in the 1990/1991 AMI Validation Study Data Set ($N = 971$ cases)

		Risk of Mortality Reabstracted ICD-9-CM Codes		*Risk of Mortality Reabstracted ICD-9-CM Codes Present at Admission*	
Model Components		*Alone*	*+ Age/Sex*	*Alone*	*+ Age/Sex*
Risk of Mortality Original ICD-9-CM Codes	Alone	.79	—	.65	—
	+ Age/Sex	—	.80	—	.69
Risk of Mortality Reabstracted ICD-9-CM Codes	Alone			.69	—
	+ Age/Sex			—	.70
		Severity of Illness Reabstracted ICD-9-CM Codes		*Severity of Illness Reabstracted ICD-9-CM Codes Present at Admission*	
Severity of Illness Original ICD-9-CM Codes	Alone	.74	—	.67	—
	+ Age/Sex	—	.78	—	.72
Severity of Illness Reabstracted ICD-9-CM Codes	Alone			.80	—
	+ Age/Sex			—	.83

Table 4: Weighted Pearson Correlations Between Hospital-Level Expected Mortality Rates from Risk-Adjustment Models with APR-DRGs, Using All Cases in the 1990/1991 AMI Validation Study Data Set ($N = 30$ hospitals)

		Risk of Mortality Reabstracted ICD-9-CM Codes		*Risk of Mortality Reabstracted ICD-9-CM Codes Present at Admission*	
Model Components		*Alone*	*+ Age/Sex*	*Alone*	*+ Age/Sex*
Risk of Mortality	Alone	.85	—	.85	—
Original ICD-9-CM Codes	+ Age/Sex	—	.83	—	.87
Risk of Mortality	Alone			.77	—
Reabstracted ICD-9-CM Codes	+ Age/Sex			—	.72
		Severity of Illness Reabstracted ICD-9-CM Codes		*Severity of Illness Reabstracted ICD-9-CM Codes Present at Admission*	
Severity of Illness	Alone	.72	—	.74	—
Original ICD-9-CM Codes	+ Age/Sex	—	.79	—	.80
Severity of Illness	Alone			.81	—
Reabstracted ICD-9-CM Codes	+ Age/Sex			—	.87

reabstracted ICD-9-CM diagnoses instead of the diagnoses originally reported to OSHPD had a modest effect on hospital-level expected mortality rates ($r = .83$–$.85$) based on the ROM class, but their use had more effect on rates based on the SOI class ($r = .72$–$.79$). Conversely, using diagnoses present at admission instead of all reabstracted diagnoses had more effect on hospital-level expected mortality rates based on the ROM class ($r = .72$–$.77$) than it did on rates based on the SOI class ($r = 0.81$–0.87).

Effect on Hospital Classification

Among the 30 medium to high-volume hospitals in the 1990/1991 AMI Validation Study data set, ten were originally classified by OSHPD as having worse than expected mortality and ten as having better than expected mortality, based on a risk-adjustment model that included age, sex, race/ethnicity, insurance, type of admission, site of infarction, 19 ICD-9-CM comorbidities, and eight two-way interactions (Romano et al. 1997). Using APR-DRG ROM

classes alone, three hospitals were classified as worse than expected and six as better than expected. Using SOI classes alone, four hospitals were classified as worse than expected (one of which was so labeled using ROM classes) and seven as better than expected (six of which were so labeled using ROM classes). Adding age and sex to the model based on ROM classes led two additional hospitals to be classified as worse than expected and three additional hospitals to be classified as better than expected (replacing two hospitals that lost that label). The addition of age and sex to the model based on SOI classes did not alter the list of hospitals classified as worse than expected, but it led two additional hospitals to be classified as better than expected.

Blinded reabstraction of medical records by specially trained coders led to more notable changes in the classification of hospitals using APR-DRG ROM classes (Table 5). Using only diagnoses reabstracted from prehospital, emergency room, and admission notes ("present at admission") to assign APR-DRG ROM classes led to further classification changes. Of the 30 hospitals, nine were classified as worse than expected using any of the three sets of ICD-9-CM diagnoses described in Table 5, but none was so classified using all sets. Eight hospitals were classified as better than expected using any of these three sets of diagnoses, but only three hospitals were so classified using all sets. Reliability statistics (Table 5) confirm that the use of hospital-reported ICD-9-CM diagnoses, all reabstracted diagnoses, and reabstracted diagnoses present at admission to assign the ROM class generated only "fair" agreement in classifying hospital performance (Landis and Koch 1977). By contrast, agreement in classifying hospital performance was "moderate" when the same three sets of diagnoses were used to assign the SOI class.

CONCLUSIONS

Through use of a large sample of AMI cases from nonfederal acute care hospitals in California, we found that the APR-DRG ROM class was a strong predictor of death and that it had better discrimination than did the SOI class that had been previously evaluated (Iezzoni, Ash, Shwartz, et al. 1996). However, the excellent performance of this measure was largely attributable to its inclusion of both comorbidities and complications. When conditions diagnosed after admission were not used to assign APR-DRGs, the predictive performance of both ROM and SOI classes fell. Indeed, the SOI class, which was not designed to predict mortality, emerged as the preferred

Table 5: Impact of Data Source on the Classification of Hospitals as Risk-Adjusted Mortality Outliers Using APR-DRG ROM Classes for All Cases in the 1990/1991 AMI Validation Study Data Set ($N = 30$ hospitals)

	Hospital Classification Based on Reabstracted ICD-9-CM Diagnoses		
	Mortality Higher than Expected	*Mortality Neither Higher nor Lower*	*Mortality Lower than Expected*
Hospital Classification Based on Original ICD-9-CM Diagnoses			
Mortality higher than expected	1	2	0
Mortality neither higher nor lower	2	18	1
Mortality lower than expected	0	3	3
Hospital Classification Based on Reabstracted ICD-9-CM Diagnoses Present at Admission			
Mortality higher than expected	1	5	0
Mortality neither higher nor lower	2	15	0
Mortality lower than expected	0	3	4

Note: Reliability statistics for ROM classes: $\kappa = 0.52$ better than expected, $\kappa = 0.26$ worse than expected, $\kappa = 0.38$ overall comparing hospital-reported with all reabstracted diagnoses; $\kappa = 0.67$ better than expected, $\kappa = 0.10$ worse than expected, $\kappa = 0.35$ overall comparing all reabstracted diagnoses with reabstracted diagnoses present at admission. Reliability statistics for SOI classes (data not shown): $\kappa = 0.56$ better than expected, $\kappa = 0.71$ worse than expected, $\kappa = 0.57$ overall comparing hospital-reported with all reabstracted diagnoses; $\kappa = 0.63$ better than expected, $\kappa = 0.35$ worse than expected; $\kappa = 0.45$ overall comparing all reabstracted diagnoses with reabstracted diagnoses present at admission.

risk-adjustment tool in this analysis. Other researchers have argued that the apparent superiority of measures based on administrative data may reflect an adjustment for complications that develop after admission (Pine et al. 1997; Iezzoni 1997a,b), but this study is the first to confirm the hypothesis using an off-the-shelf system.

The decrease in predictive performance that results from using only diagnoses "present at admission" to assign APR-DRGs has a major impact on the classification of hospitals based on risk-adjusted mortality. Among the 30 hospitals included in OSHPD's AMI Validation Study, agreement was only fair in classifying hospital mortality as significantly better than expected, significantly worse than expected, or neither of the above (using the ROM class with or without demographic variables).

Several other points about the predictive performance of APR-DRGs deserve mention. First, demographic variables significantly improved any risk

adjustment based on APR-DRGs. This indicated that APR-DRGs do not fully capture the effects of age and sex on mortality among AMI patients; it is not surprising, because all adults are classified as less than 70, or 70 or more, years of age when the APR-DRG ROM score is assigned (3M Health Information Systems 1995). The effects of age and sex could be modeled almost equally well using either dummy cross-classification variables or linear piecewise and interaction effects. Unfortunately, all models that included age and sex had poor calibration; they overestimated the risk of death among the 30 percent of patients at lowest risk and the 3–5 percent of patients at highest risk. This problem is more likely related to the omission of an interaction term (e.g., between age or sex and ROM class) than to a lack of linearity in the logit function.

Second, the information contained in SOI classification is not fully captured by ROM classification; the addition of SOI classes to a model containing ROM classes would further enhance discrimination. However, this improvement is relatively small and it may have little practical significance.

Third, the APR-DRG ROM score would have even better predictive power if hospitals reported all of the ICD-9-CM diagnoses documented in the medical record. The *c*-statistics of .920–.929 that we obtained using reabstracted data are higher than those reported in several studies that used detailed clinical data (Pine et al. 1997; Iezzoni, Ash, Shwartz, et al. 1996; Pennsylvania Health Care Cost Containment Council 1996; Alemi, Rice, and Hankins 1990; Office of Statewide Health Planning and Development 1996). As shown in Table 2, these parameters are misleading because both comorbidities and potentially preventable complications are used to assign ROM classes.

These findings improve our understanding of the ways in which different risk-adjustment methods should be used. The APR-DRG ROM system, which was the focus of our study, is a powerful risk-adjustment tool for inpatient mortality. However, it achieves such high performance by including all clinically relevant diagnoses assigned to a patient, including diagnoses that developed after admission. Some of these late diagnoses probably reflect the natural history of a patient's disease (e.g., third-degree heart block), which healthcare providers may be powerless to alter. However, many of these late diagnoses may be the result of poor medical care (e.g., congestive heart failure attributable to volume overload or nosocomial infections traceable to poor hand-washing practices). In practice, these two categories of late diagnoses are probably impossible to distinguish.

As a result, APR-DRGs are a suitable risk-adjustment tool only if one wishes to adjust for both comorbidities and complications. Such a strategy

might be appropriate if one is interested in establishing whether or not hospitals and physicians rescue patients who experience complications (Silber et al. 1992). It may also be appropriate for a multihospital system that aims to compare performance at its own facilities. However, we believe, along with others (e.g., Selker, Griffith, and D'Agostino 1991), that it is inappropriate to adjust for potentially preventable complications in public report cards on provider performance. A risk-adjustment tool that is less susceptible to the effects of complications should be selected for this purpose. Unfortunately, no published data are available to compare the performance of commercial risk-adjustment products on this dimension. The current version of APR-DRGs (Version 15) may be significantly better than the version we evaluated because most "complications of surgical and medical care" (ICD-9-CM 996–999) are no longer used in risk-adjustment. Two states (California and New York) now require hospitals to report whether or not each ICD-9-CM diagnosis was present at admission. With this information, future analysts using APR-DRGs or other risk-adjustment products may be better able to separate the effects of severity of illness at admission and patients' subsequent clinical trajectories.

NOTES

1. For more information, see http://www.fdhc.state.fl.us/schs/hospguide/hospguide.htm, http://www.mha.org/mhr4/, and http://hlunix.hl.state.ut.us/hda/.
2. For more information, see the California Hospital Outcomes Project Web site at http://www.oshpd.cahwnet.gov/hpp/chop.html.
3. Transfers, prior admissions, and readmissions were linked if they matched exactly on the encrypted SSN and at least two of the three elements of birth date. Transfer records were also linked if they matched exactly on date of birth, gender, and zip code of residence and had admission and discharge dates consistent with an interhospital transfer.
4. These criteria were designed by our panel of consulting cardiologists to exclude patients who might not actually have suffered an AMI. For example, patients with very short stays (e.g., less than three days in 1990/1991, less than two days in 1991–1993) probably ruled out for AMI but received an AMI code because no alternative diagnosis was documented by the physician. Patients with a recent AMI probably experienced extension of their prior infarct. Patients involved in a transport accident may have suffered a myocardial contusion, with resulting elevation of cardiac enzymes. Severe aortic stenosis may be associated with poor coronary flow, but this criterion was dropped after the Validation Study because most cases did not appear to be clinically significant.

5. This cutoff was selected to yield a 5 percent chance of improperly excluding one or more of the 431 eligible hospitals, correcting for multiple comparisons (four per hospital).
6. This exclusion was justified by our use of linked hospital discharge abstracts to determine whether each patient was dead or alive 30 days after admission. It was unnecessary in the more recent data set because discharge abstracts were linked with death certificates, as described later in the text.
7. This linkage was accomplished using either a hard match on SSN and gender with a soft match on date of birth, or a hard match on gender, race, five-digit zip code, and date of birth with a soft match on SSN. Details are available on request.

REFERENCES

Alemi, F., J. Rice, and R. Hankins. 1990. "Predicting In-Hospital Survival of Myocardial Infarction." *Medical Care* 28 (9): 762–75.

Averill, R. F., J. H. Muldoon, J. C. Vertrees, N. I. Goldfield, R. L. Mullin, E. C. Fineran, M. Z. Zhang, B. Steinbeck, and T. Grant. 1997. "The Evolution of Casemix Measurement Using Diagnosis Related Groups (DRGs)." 3M HIS Working Paper. Wallinford, CT: 3M Health Information Systems.

Brailer, D. J., E. Kroch, M. V. Pauly, and J. Huang. 1996. "Comorbidity-adjusted Complication Risk. A New Outcome Quality Measure." *Medical Care* 34 (5): 490–505.

Cameron, A. C., and F. A. G. Windmeijer. 1997. "An R-Squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models." *Journal of Econometrics* 77 (2): 329–42.

Hanley, J. A., and B. J. McNeil. 1982. "The Meaning and Use of the Area Under a Receiver Operating Characteristic Curve." *Radiology* 143 (1): 29–36.

Hosmer, D. W. Jr., and S. Lemeshow. 1989. *Applied Logistic Regression.* New York: John Wiley & Sons.

Hosmer, D. W., T. Hosmer, S. LeCessie, and S. Lemeshow. 1997. "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model." *Statistics in Medicine* 16 (9): 965–80.

Iezzoni, L. I., J. Daley, T. Heeren, S. M. Foley, E. S. Fisher, C. Duncan, J. S. Hughes, and G. A. Coffman. 1994. "Identifying Complications of Care Using Administrative Data." *Medical Care* 32 (7): 700–15.

Iezzoni, L. I., M. Shwartz, A. S. Ash, J. S. Hughes, J. Daley, and Y. D. Mackiernan. 1995. "Using Severity-adjusted Stroke Mortality Rates to Judge Hospitals." *International Journal for Quality in Health Care* 7 (2): 81–94.

Iezzoni, L. I., A. S. Ash, M. Shwartz, J. Daley, J. S. Hughes, and Y. D. Mackiernan. 1996. "Judging Hospitals by Severity-adjusted Mortality Rates: The Influence of the Severity-Adjustment Method." *American Journal of Public Health* 86 (10): 1379–87.

Iezzoni, L. I., M. Shwartz, A. S. Ash, J. S. Hughes, J. Daley, and Y. D. Mackiernan. 1996. "Severity Measurement Methods and Judging Hospital Death Rates for Pneumonia." *Medical Care* 34 (1): 11–28.

Iezzoni, L. I. 1997a. "The Risks of Risk Adjustment." JAMA: *Journal of the American Medical Association* 278 (19): 1600–607.

———. 1997b. "Risk Adjustment and Current Health Policy Initiatives." In *Risk Adjustment for Measuring Healthcare Outcomes,* edited by L. I. Iezzoni, pp. 517–95. Chicago: Health Administration Press.

Iezzoni, L. I., A. S. Ash, M. Shwartz, B. E. Landon, Y. D. Mackiernan. 1998. "Predicting In-Hospital Deaths from CABG Surgery." *Medical Care* 36 (1): 28–39.

Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–74.

Levy, P. S. 1991. *Sampling of Populations: Methods and Applications.* Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons.

New York State Department of Health. 1996. *Coronary Artery Bypass Surgery in New York State, 1993–1995.* Albany, NY.

Office of Statewide Health Planning and Development. 1996. *Second Report of the California Hospital Outcomes Project: Acute Myocardial Infarction. Volume Two: Technical Appendix.* Sacramento, CA.

———. 1993. *Annual Report of the California Hospital Outcomes Project. Volume Two: Technical Appendix.* Sacramento, CA.

Pennsylvania Health Care Cost Containment Council. 1996. *Focus on Heart Attack in Pennsylvania: Research Methods and Results.* Harrisburg, PA.

Pine M., M. Norusis, B. Jones, and G. E. Rosenthal. 1997. "Predictions of Hospital Mortality Rates: A Comparison of Data Sources." *Annals of Internal Medicine* 126 (5): 347–54.

Romano, P. S., H. S. Luft, J. A. Rainwater, and A. P. Zach. 1997. *Report on Heart Attack 1991–1993, Volume Two: Technical Guide.* Sacramento, CA: California Office of Statewide Health Planning and Development.

Selker, H. P., J. L. Griffith, and R. B. D'Agostino. 1991. "A Time-Insensitive Predictive Instrument for Acute Myocardial Infarction Mortality: A Multicenter Study." *Medical Care* 29 (12): 1196–211.

Silber, J. H., S. V. Williams, H. Krakauer, and J. S. Schwartz. 1992. "Hospital and Patient Characteristics Associated with Death After Surgery. A Study of Adverse Occurrence and Failure to Rescue." *Medical Care* 30 (7): 615–29.

Soeken, K. L., and P. A. Prescott. 1986. "Issues in the Use of Kappa to Estimate Reliability." *Medical Care* 24 (8): 733–41.

3M Health Information Systems. 1995. *All Patient Refined Diagnosis Related Groups (APR-DRGs), Version 12.0. Definitions Manual.* Wallingford, CT.