RAND

Assessing the Performance of Mortality Prediction Models

David C. Hadorn, Emmett B. Keeler, William H. Rogers, Robert H. Brook

RAND/UCLA/Harvard Center for Health Care Financing Policy Research

The research described in this report was sponsored by the Health Care Financing Administration, U.S. Department of Health and Human Services, under Cooperative Agreement No. 99-C-98489/9-08.

ISBN: 0-8330-1335-1

Published 1993 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
To obtain information about RAND studies or to order documents,
call Customer Service, (310) 393-0411, extension 6686

RAND

Assessing the Performance of Mortality Prediction Models

David C. Hadorn, Emmett B. Keeler, William H. Rogers, Robert H. Brook

Prepared for the
Health Care Financing Administration,
U.S. Department of Health and Human Services

RAND/UCLA/Harvard Center for Health Care Financing Policy Research

Preface

The skyrocketing costs of health care, together with persistent concerns about the provision of inappropriate care, has fostered within the United States a broad-ranging effort to determine the clinical outcomes of medical and surgical services. However, despite more than two decades of work at RAND and elsewhere to conceptualize and measure health status, outcome assessment remains a relatively primitive science. Most efforts to date have been limited to measurement of mortality rates—one of the few "hard" outcomes available for study.

Even the assessment of mortality rates is tricky, however. In particular, before it can be determined whether an observed mortality rate is good or bad, the "expected" mortality rate for that particular sample of patients must be specified. This expectation, in turn, must be derived from a consideration of various clinical and demographic factors that raise or lower the risk of death.

The work described in this report was performed in an effort to advance the science of mortality prediction by (1) pointing out the need for standards in the development, testing, and application of mortality prediction models and (2) beginning the process of sketching out what those standards might look like. Toward these ends, the authors surveyed existing published literature on (certain types of) mortality prediction in medicine, performed some empirical work related to the measurement of prediction model performance, and convened an expert panel to discuss the issue of standards in the field. This report describes the results of those efforts.

This work was performed under Cooperative Agreement Number 99-C-98489/9-08 with the Health Care Financing Administration.

Contents

Figure

Tables

Summary

Observed health care outcomes are increasingly relied upon for evaluating the appropriateness and quality of medical care. But outcome assessment requires estimation of what *should* have happened given proper (or optimal) care. This requirement has resulted in the creation of a host of statistically based outcome prediction models and systems, often referred to as "risk-adjustment" or "case-mix" measures. These systems are used to adjust observed mortality rates for differences in patients' severity of illness.

We began this project with the mission of evaluating these various measures and determining the extent to which performance comparisons among measures was feasible. We quickly realized that these were impossible tasks because competing systems are developed and evaluated on fundamentally different data bases. Moreover, the systems were evaluated using non-comparable statistics.

Any attempt to declare one (or a few) mortality prediction models "winners" would also be hampered by the rapidity with which systems evolve. The specific systems described in published literature often bear little resemblence to the same systems in existence today; moreover, worthy new competitors can appear at any time.

For all of the foregoing reasons, we instead concentrated on developing a set of criteria that vendors of risk-assessment systems should meet in order to permit consumers to intelligently evaluate them in the future. This report describes our ideas about what those criteria should be, as well as the opinions expressed at a meeting of experts that was convened as part of this project.

We first describe the background in which health outcome prediction (i.e., risk-adjustment or case-mix measurement) has become important. We then discuss the documentation (or lack thereof) available for several well-known risk assessment systems. The next section of the report describes an empirical exercise and theoretical example showing that the nature of the evaluation dataset strongly influences the evaluation it produces. The results of the expert meeting are then described in detail. The report concludes with our recommendations concerning desirable evaluation criteria.

Acknowledgments

We wish to acknowledge the substantial contributions to this project of the experts who were kind enough to spend a day in Washington to discuss the desirability of standards in the field of mortality prediction in medicine. These individuals are listed in Appendix A. We offer special thanks to Arlene Ash, Ph.D., Mark Blumberg, M.D., Lisa Iezzoni, M.D., M.Sc., Patrick Romano, M.D., and Harry Selker, M.D., MSPH, for their very useful advice and assistance. Many other individuals also offered helpful suggestions during the course of this project, including Farrokh Alemi, Ph.D., Susan DesHarnais, Ph.D., and Jesse Green, Ph.D. We also appreciate the support of our Project Officer, Harry Savitt, Ph.D., and of Steven Jencks, M.D., who also was instrumental in assisting with the conceptualization of this project. Finally, we acknowlege the assistance of Carole Oken, M.A., in preparing Table 2. Although we are indebted to all these individuals, none of them necessarily agrees with any of the conclusions or recommendations contained in this report.

1. Introduction

Mounting concerns about the cost and quality of health care have resulted in a multi-faceted national effort to determine the health outcomes associated with medical and surgical services. Outcomes research will lead, it is hoped, to data-based recommendations concerning appropriate clinical management strategies, which in turn will be encapsulated in clinical practice guidelines.

Beyond the evaluation of health services, outcomes research is also potentially useful for assessing the quality of care administered by hospitals and physicians. Here, observed outcomes of care are compared with "expected outcomes," i.e., those outcomes that most patients (or the "average patient") with the same set of relevant characteristics should (or does) experience. Specification of expected outcomes is a complicated task, of course, not only because outcomes other than mortality are difficult to define, but also because of the probabilistic nature of health outcomes. For example, even the best candidate for a particular operation may die on the operating room table, through no fault of the surgeon, and even a poor candidate for some chemotherapy regimen may obtain a complete remission.

In effect, specification of expected outcomes represents *predictions* about what "ought" to happen (or to have happened) to a particular patient, or group of patients, given application of good (or average) quality care. Predictions are based on relevant prognostic factors, such as the patient's age or the extent of disease or physiological disruption. These factors are the same ones that physicians and other health care providers have relied on for centuries to answer questions such as, "How long have I got, doc?" and "What are the chances it's cancer?" Indeed, the art of prognostication has always played a vital role in clinical practice.

Stimulated by rapidly rising costs and continued concerns about variations in the content and quality of health care, the science of clinical prediction is steadily evolving. Dozens of statistical models have been developed to predict mortality rates, length of stay, resource use (e.g., charges), and readmission or complication rates (Wasson et al., 1985). In addition, hundreds of models have been developed to "predict" the presence of a disease or condition, i.e., to diagnose. Statistical prediction models rely on the same clinical and demographic factors (e.g., age, blood pressure) used by clinicians to arrive at prognostic judgments. Unlike clinicians, however, models assign explicit weights to these factors based on their

observed statistical association with the outcome of interest (e.g., mortality) in some sample of patients. As a result, prediction models render precise (if not always accurate) predictions of outcome or diagnosis.

The most significant area within the new field of clinical prediction science concerns the assessment of patient mortality risk. Mortality prediction models generate explicit probabilities of death during hospitalization or within a specified period of time (often 30 days) after admission, discharge, or surgery. The estimated probabilities of death reflect measurable differences in the severity of illness of patients across physicians and hospitals. Observed departures from expected mortality rates are then considered (rightly or wrongly) to be evidence of better- or worse-than-average quality of care. The yearly release by the Health Care Financing Administration (HCFA) of hospital mortality statistics represents the most prominent example of this process.

Severity-adjusted mortality rates are increasingly used to draw inferences about hospital and physician quality. Besides the yearly HCFA release of hospital mortality rates, several other large-scale quality assessment projects that use differences between expected and observed mortality as their principal "measure of quality" are either completed or in progress (e.g., New York Cardiac Surgery Reporting System, documented in Hannan et al., 1990, and the Cleveland Coalition, documented in Meyer, 1990). At least one payor, the Minnesota Blue Cross/Blue Shield, has developed a policy linking reimbursement to severity-adjusted outcomes ("Minnesota Blues' Payment Plan...," 1991). Mortality prediction models are potentially attractive for an increasingly broad range of applications within the health care system, as summarized in Table 1. Many of these applications are ethically and politically very sensitive, and have potentially quite significant consequences.

Are the mortality prediction models available today ready to be used for these tasks? Are they ready for some purposes, but not for others? In seeking to answer these questions, many factors are relevant, including the cost and feasibility of implementing prediction systems within hospitals or other institutions, and the extent to which the clinical variables used in the model are resistant to manipulation and subjectivity.

More fundamentally, mortality prediction models should be both reliable and valid before they are used for any of the applications listed in Table 1. In this context, reliability and validity each have two different levels of meaning: (1) the level of the data elements, or independent variables, and (2) the level of the nu-

Table 1

Current and Potential Uses of Outcome Prediction Models

Institutional Level

- Educational use in institutional quality improvement efforts, such as a community hospital comparing its outcomes against national averages and ranges. Internal use only.
- Institutional self-monitoring for competitive or contractual reasons, to attract patients and insured populations. Hospitals have been advertising their comparative outcomes for several years; this trend is increasing.
- Formal monitoring by regulatory agencies and payors. These data would be made available to the public and potentially used as the basis for administrative sanctions or for determination of hospitals' bond ratings. (There are signs that this is already happening; Nemes, 1991.)
- Used to adjust the outcomes observed in clinical trials across treatment arms (as an alternative to excluding from the trials very ill patients or patients with significant co-morbidity).
- Patients could use outcome data to select among different hospitals or other institutions

Physician Level

- Educational use for internal clinical quality improvement efforts and related activities. Individual physicians could compare their patients' outcomes with local, regional, and national averages and ranges. (Note sample size requirements for this application, addressed below.)
- Institutions, including hospitals and managed care organizations, could use outcome information on individual physicians for internal quality review and for decisions about whether to grant, extend, or revoke practice privileges.
- Formal monitoring of outcomes could be made the basis for public disclosure about physician quality of care. Presumably, patients would prefer to seek the services of physicians with low patient mortality.
- In extreme cases, sustained patterns of poor (severity-adjusted) outcomes could be used as the basis for formal sanctions or license withdrawal.

Patient Level

- Patients could be told their estimated numerical probability of survival as part of the counseling and decisionmaking process. This information would be presented, together with additional factors relevant to making treatment decisions, e.g., whether to undergo surgery or to be admitted to intensive care.
- Resource allocation purposes, including triage in the setting of an overcrowded intensive care unit.

merical probability assessment itself. The first level of reliability (or inter-rater reliability) describes whether different people collecting data at different times or in different locations obtain the same value for a specific independent variable in a specific patient. Variables that are subjective, difficult to interpret, or difficult

to define may have poor reliability. Many developers of mortality prediction models have assessed this level of model reliability (see Table 2).

The second level of reliability refers to whether different users of the model arrive at the same (or similar) predictions of mortality risk for the same (or similar) patients. This level of reliability takes into account the entire process of data acquisition, entry of data into the model, and probability calculations. Few (if any) model developers have assessed this level of model reliability.

By contrast, validity refers to whether the instrument is actually measuring the underlying concept of interest, i.e., probability of death. The first level of model validity refers to whether the variables included in the model actually correspond to increased or decreased mortality risk. For example, five different abstractors may all agree that the serum bilirubin is listed in a patient's chart. However, that independent variable may be invalid if serum bilirubin does not actually contribute to mortality risk. Even more critical, of course, is the validity of the prediction itself—whether the numerical probability assigned to an individual patient or group of patients actually corresponds to the risk of mortality. In common statistical parlance, this level of validity is often referred to as model accuracy or performance.

The present project focused on this second level of model validity, or accuracy. The question of model accuracy really has two parts: (1) how accurate are current models? (or what is the range of accuracy across models?) and (2) how accurate *should* models be before they are used for various purposes? The second question requires that value judgments be made about the outcomes of model use, something that remains to be done. The first question, on the other hand, should be amenable to relatively objective, scientific scrutiny and would entail the specification of evaluation criteria.

With the above as preamble, then, we set out to compare published information about the predictive performance of several selected mortality prediction models.

2. Survey of Existing Mortality Prediction Models

We surveyed existing models to determine how model performance is currently being reported, as well as to learn the general range of performance reported using these statistics. We limited our review to peer-reviewed literature dealing with models that were designed to predict mortality after hospitalization for one or more of the major mortality-producing conditions affecting the Medicare population (e.g., acute myocardial infarction, stroke, pneumonia, and congestive heart failure) or among general medical or medical intensive care unit patients. Models designed to predict death from trauma, or that applied to pediatric patients or to other specific diseases or conditions, were excluded. Models used for "severity adjustment" or "risk adjustment" were included in our review when they were in effect mortality prediction models designed to apply to the medical populations described above.

Model developers use a wide variety of statistics and techniques to measure the performance (i.e., the accuracy or validity) of their models. Moreover, sample characteristics vary considerably across studies, as do the techniques used to develop the models. Table 2 summarizes these findings. At most one study per research group was included in the table to minimize methodological redundancy. Our purpose in summarizing these studies is not to provide an exhaustive list, but rather to illustrate the ranges of studies and the types of test statistics currently employed in the field of mortality prediction.

We found that the two most commonly reported statistics are R^2 and the area under the receiver-operating characteristic (ROC) curve, or c-index. R^2 can be interpreted as the proportion of outcome variance explained, or "accounted for," by a model. The theoretical upper limit of R^2 is 1.0, but the practical limit for models that predict dichotomous outcome variables (e.g., alive or dead within 30 days) is substantially lower. R^2 statistics reported for models predicting death in the hospital or within 30 days ranged from about 0.045 to 0.388 in the studies we reviewed. (Note that this applies to performance using individual patients as the unit of analysis. Model performance is substantially higher when groups of patients (e.g., hospitals) are the unit of analysis (DesHarnais et al., 1988). This phenomenon is related to the so-called ecological bias (Greenfield and Morgenstern, 1989). The highest values of R^2 reported for most diseases and conditions ranged

Table 2

Mortality Prediction Models

AUTHOR/DATE	NEW MODEL OR TEST OF EXISTING MODEL	# & TYPE OF SAMPLE	TYPE OF INDEPENDENT VARIABLES	METHOD USED TO DEVELOP MODEL	RANGE OF PEFORMANCE LEVELS								MORTALITY RATE	COMMENTS
Alemi, et al., 1990	Test of 7 existing systems.	775 pts. w/acute MI w/medical & surgical treatment performance. Retested on subsample of 355 pts. w/medical treatment only.	Varies by individual system.	N/A	SYSTEM	1	2	3	4	5	6**	7	22% In-hospital	Coding reliability > 99%. 4 of 7 symptoms coded by vendor personnel -- "black box" software.
					ROC	0.70	0.69	0.44	0.74	0.66	N/A	0.73		
					*Chi² GOODNESS OF FIT	67	75	14	125	36	N/A	162		
					* All significant p < .001 ** Tested only in subsample									
Blumberg, 1991	Test of existing system.	3037 Medicare pts with acute MI.	5 level ordinal scale using 260 clinical variables.	Clinical judgment.	Standardized mortality ratios and associated chi² across different levels of additional predictors. chi² (df) p Age 34.0(4) < .001 AMI location 47.9(2) < .001 CHF 5.9(3) < .150 ECG & enzyme tests 29.(5) < .001								24.2% 30-day	Concluded that system produced potentially biased mortality estimates. Coding accuracy not assessed directly (PRO abstractors trained by vendor to >95% reliability.)
Charlson, et al., 1986	New	604 unselected medical admits.	Ordinal rating by resident physicians of severity, co-morbidity, stability, functional status, admit reason seriousness.	Logistic, OLS, Cox regression.	Chi² 69.2(5df) p<0.001 for severity only. Incremental predictive improvements by adding variables (e.g. functional status).								10.9% In-hospital	Did not compare performance against regression based model on same sample; rating reliability not examined; coding accuracy not assessed. Cross-validated in subsequent study (Charlson, et al., 1987).

Table 2—Continued

AUTHOR/DATE	NEW MODEL OR TEST OF EXISTING MODEL	# & TYPE OF SAMPLE	TYPE OF INDEPENDENT VARIABLES	METHOD USED TO DEVELOP MODEL	RANGE OF PERFORMANCE VALUES	MORTALITY RATE	COMMENTS
Daley, et al., 1988	New	5,888 hospitalized Medicare pts. with 4 diseases (CHF, stroke, pneumonia, acute MI) analysis stratified by disease.	APACHE plus additional clinical and lab variables from 1st 24h, from medical record.	Logistic regression goodness of fit using chi² analysis.	STROKE / PNEUMONIA / MYOCARDIAL INFARCTION / CONGESTIVE HEART FAILURE Medicare Mortality Prediction System: .253 (.187) / .181 (.162) / .138 (.223) / .148 (.147) APACHE II: .220 (.187) / .141 (.162) / .092 (.223) / .046 (.147) Values are internally cross-validated R^2. (30-day mortality rates.)		Reliability testing kappa generally between 0.79 - 0.88. Model developed in random 2/3 of sample, tested in remaining 1/3.
DesHarnais, et al., 1988	New	300 hospitals, 2 cohorts, all deaths 1983, and all deaths 1984, CPHA database.	Claims data including 1° and 2° diagnoses, comorbidity age and procedures.	2x3 contingency table (death rate <5%) Logistic regression (death rate >5%) DRG clusters unit of analysis.	Correlations btwn expected and observed mortality (R): 1983 / 1984 Hospital Level: 0.84 / 0.81 DRG-Cluster Level: 0.86 / 0.85 DRG-Cluster by Hosp.: 0.86 / 0.85 Also, model reduced prediction errors per cluster by 14% vs predicting survival base rate.	2.73%-3.05% In hospital	Coding accuracy not reported.
Detre, et al., 1981	New	508 male pts. From VA coronary artery surgery study to develop function. 686 pts. to test function.	4 dichotomous predictors: NYHA Class < III vs ≥ Class III, H/O Hypertension, H/O MI, ST-segment depression.	Cox-Breslow life table regression for risk function, probability of dying to create risk terciles. Jackknife procedure to validate risk function.	Mantell-Hanszel Chi-square for risk function results. High-risk Chi2 18.8 $p<.0005$. Low-risk Chi2 = 7.64, $p>0.5$. Regression slope-observed on predicted = 0.829 ± 0.155.	5 years, calculated by terciles only.	Cross-val with 535 pts. at a different institution. "validity regression slope" 0.728 ± 0.1444. Coding accuracy not addressed but performed during VA study.
Green, et al., 1990	Test of existing system.	34,352 Medicare pts. With cancer, severe acute heart disease, stroke, pulmonary disease, low risk heart disease. Compared with HCFA mortality prediction model alone.	History, physical, lab data from medical record.	Clinical judgment.	R^2*: HCFA ALONE 2.5% / HCFA + SEVERITY 21.5% (average) for 5 diseases (* Weighted average,) C(statistic): 0.60-0.64 / 0.72-0.84 STROKE / PNEUMONIA / HEART FAILURE Simulated HCFA Model: .051 (.19) / 0.19 (.09) / .039 (.27) Simulated HCFA Model Plus Severity of Illness: .298 (.19) / .173 (.09) / .256 (.27) Values are internally cross-validated R^2. (30-day mortality rates.)		Coding accuracy > 90%.

Table 2—Continued

AUTHOR/DATE	NEW MODEL OR TEST OF EXISTING MODEL	# & TYPE OF SAMPLE	TYPE OF INDEPENDENT VARIABLES	METHOD USED TO DEVELOP MODEL	RANGE OF PERFORMANCE VALUES	MORTALITY RATE	COMMENTS
Horn, et al., 1991	Test of existing system	2378 pts. in 27 high-volume DRGs.	Four level ordinal scale based on medical record data.	Clinical judgment.	R=0.603, range 0.469 to 0.682 for individual DRGs. C=0.891, range 0.833 to 0.975 for individual DRGs.	In-hospital 3-7% overall. Range 0.60% (level 1)-49.5% (level 4).	Kappa for coding 0.88-0.90
Iezzoni, et al., 1991	Test of existing system	20,985 randomly selected Medicare admissions with one of six conditions: acute MI, pneumonia OR chronic obstructive pulmonary disease, coronary artery revascularization, congestive heart failure, cholecystectomy; and prostate surgery.	Stratified into one of five overall severity levels based on clinical indicators.	Clinical judgment.	R2 calculated both for admission and "mid-stay review" (approximately one week into hospitalization). C-index calculated but not reported "since no substantive differences emerged" from looking at this statistic. STROKE: SIMULATED HCFA MODEL .051 (.19) PNEUMONIA: .019 (.09) HEART FAILURE: .039 (.27) Values are internally cross-validated R^2, (30-day mortality rates).		Authors acknowledge vulnerability of model to overfitting. Number of patients in each cell of the admission x mid-stay score matrix not stated.
Keeler, et al., 1990	New	14,012 Medicare pts. with CHF, AMI, cerebrovascular accident, pneumonia, hip fracture.	Sickness at admission model using a range of 18-25 variables including 13 APACHE II variables from medical record. From 66-80 initially tried.	Stepwise logistic regression.	RAND/PPS: STROKE .30 (.21); PNEUMONIA .26 (.16); MYOCARDIAL INFARCTION 22.(.24); HEART FAILURE .12 (.14) Values are cross-validated R^2, (30-day mortality rates.)		Coding accuracy extensively assessed at item level (see RAND Report R3931-HCFA on PPS study especially Table 2-6). Model developed in random 2/3 subsample tested in remaining 1/3.
Knaus, et al., 1985	New (revised)	5,815 unselected ICU admissions.	History, physical and laboratory findings from medical record.	Clinical judgment.	SYSTEM: APACHE II R^2 0.319, ROC 0.863, chi^2 1634.5 P<.001; APACHE I R^2 0.310, ROC 0.851, chi^2 1537.1 P<.001	20% In-hospital	Interobserver reliability for independent variables = 96%

Table 2—Continued

AUTHOR/DATE	NEW MODEL OR TEST OF EXISTING MODEL	# & TYPE OF SAMPLE	TYPE OF INDEPENDENT VARIABLES	METHOD USED TO DEVELOP MODEL	RANGE OF PERFORMANCE VALUES	MORTALITY RATE	COMMENTS
Lemeshow, et al., 1985	New	755 consecutive ICU pts.	History, physical and lab values from medical record.	Linear discriminate function analysis with forward stepping.	Hosmer-Lemeshow Goodness of Fit statistic *Cg 6.34-6.82 (df8) p=.6092-.5562. Classification table, total correct = 85% - 87%.	19.7% In-hospital	Inter-rater reliability Kappa 0.7-1.0. Intra-rater Kappa 0.85-1.0 model. Later cross-validated in subsequent studies (eg, Teres, et al, 1987).
Selker, et al., 1991b	Logistic regression weighted according to functional class and presence of angina.	5,773 pts with chief complaint of chest pain or similar symptoms. Tested second cohort of 1,387 pts.	Demographic and clinical data from medical record.	Logistic regression weighted according to functional class and presence of angina.	1) Area under ROC curve = 0.76-0.85. 2) Hosmer-Lemeshow on deciles of predicted probabilities: rejected hypothesis of no fit (chi-square = 10.3; P = 0.24).	9.9 - 19.3% (6 hospitals)	Coding accuracy reported in previous study. Cross-validated in separate population.
Smith, et al., 1991	Compared new regressions to HCFA categories and existing system.	41,963 Medicare patients from 81 hospitals.	Administrative (HCFA categories) and clinical variables (existing system and regressions).	Logistic regression for new model.	Model Description df Chi-square R^2 Null 1 24,487 0 HCFA system 17 20,281 17.2 Existing proprietary 85 17,309 29.3 Regressions 186 14,997 38.8 Also reported indirectly standardized mortality ratio and ratio of average true positive to average false positive for each model.	14.1% 30 day	Regressions not cross-validated.
Thomas and Ashcraft, 1989	Test of 5 existing systems. Interrater reliability only.	431 cases in 11 DRG clusters.	Multivariate, scales, clinical groupings and categorical assignments, weights. Varies by system.	N/A	System 1 2 3 4 5 R_i 0.874 0.631 0.461 0.446 0.879 K 0.726 0.840 0.708 0.524 0.892 Gamma 0.925 0.920 0.736 0.596 0.922 Tau-B 0.837 0.780 0.621 0.496 0.872	N/A	Vendor staff coded one system; others coded by hospital personnel.
Weingarten, et al., 1990	Test of two existing models.	246 patients with acute stroke, including 49 over-sampled deaths.	16 clinical and physiological variables (Model 1); 3 clinical variables (Model 2).	Clinical judgment.	Correlation coefficient R of model with death = 0.50 for both models when 49 oversampled deaths included; R = 0.40 for Model 1 and 0.38 for Model 2 when oversampled deaths excluded.	9% in hospital (excluding over-sampled deaths).	None

between about 0.20 and 0.30. Thus, about one-quarter of the variance in hospital or 30-day mortality is accounted for by these models. The c-index provides an indication of a model's ability to distinguish cases that have the outcome of interest (e.g., mortality) from those that do not. This property is commonly referred to as "resolution" or "discrimination." A value of 0.50 is achieved through random predictions, or if a uniform prediction is made for everyone. The theoretical maximum of 1.00 represents perfect discrimination. C-index values for the models we reviewed typically ranged from about 0.70 to 0.80.

The observed level of performance for the models reviewed here seems adequate for assessment and comparison of severity-adjusted mortality rates, but only when sufficient numbers of patients are randomly sampled from each provider (e.g., physician, HMO, hospital). This is important in part to ensure that the spread of explanatory variables across patients is sufficiently broad to permit mortality prediction models to distinguish reliably between high- and low-risk patients. This issue is discussed in detail below.

The Problem of Chance

An equally fundamental reason why a large number of cases per provider is required before inferences about quality of care can be drawn from severity-adjusted mortality rates is that the effects of chance may swamp any true quality effects when relatively few patients are sampled. The number of cases needed to accurately specify quality effects depends on the mortality rate of the sample.[1]

Because of the effects of chance, limitations in sample size can obscure the link between quality and outcome. For example, DuBois et al. (1988) found no difference in the process of care between high- and low-outlier hospitals, although more deaths in the latter category of hospital were deemed "preventable" by a panel of physicians. Similarly, Park et al. (1990) found that most of the variance in mortality rates across hospitals could probably be accounted for by chance alone even after adjustment for severity. Data for both these studies included fewer than five patients from each hospital, so that the high- and low-outlier

[1]For example, with an overall mortality rate of 10 percent, outcome variance is (0.1 x 0.9) = 0.09. An excellent severity-adjustment measure might reduce this deviation to 0.06. The square root of this figure, 0.24, is the standard deviation of the severity-adjusted mortality rate. Dividing this latter figure by the square root of sample size produces the final standard error around mortality. Thus, a sample size of 57 (as in the RAND Prospective Payment Study) yields a final standard error of 0.24/7.5 = 0.032. This 3.2 percent standard deviation on a 10 percent mortality rate is probably too large to distinguish reliably between hospitals. A sample size of 300, on the other hand, would result in a standard error on the mean of 1.4 percent, probably an adequate level of precision for inter-hospital comparisons. Similar computations can be applied to different mortality rates.

hospitals identified by these studies may have achieved outlier status because of small sample sizes and the effects of chance.

By contrast, the RAND Prospective Payment System study collected detailed clinical information on 14,000 patients. In these data, mortality adjusted for sickness at admission strongly correlated with explicit process criteria and a structured implicit review of medical records by physicians (Kahn et al., 1990; Rubenstein et al., 1990) and differences between broad classes of hospitals could be seen (Keeler et al., 1992). Even here, however, sample size at each hospital (about 57 patients/hospital) was insufficient to draw firm conclusions about relative quality of care.

Conclusions from Literature Review

Although, as just discussed, most existing clinical-data-based mortality prediction models are probably adequate for analyses of outcomes and comparisons across providers (assuming sufficient sample sizes for each provider), the lack of standardization in performance statistics and other study characteristics (as illustrated by Table 2) effectively precludes meaningful comparison of performance across models. As noted by Iezzoni et al. (1991), "such comparisons would be hampered by differences in the data bases, predictive models, and statistical techniques used in the different studies." *Thus, it is presently not possible to recommend one model over another based on reported performance characteristics.*

Actually, as discussed below, the problem of comparing model performance runs much deeper than the simple lack of test statistic standardization. The intrinsic "predictability" of the various patient samples (as determined by the range and distribution of the samples' values on the model's independent variables) is probably the most important factor in determining how well models perform.

Despite the problem of model noncomparability, a few other tentative conclusions can be drawn from our review of existing models. *First, disease-specific models outperform models designed to predict over a wide range of patient conditions* (Daley et al., 1988; Keeler et al., 1990; Iezzoni et al., 1992; Knaus et al., 1991). There are two major explanations for this finding: (1) Specific variables can be included in disease-specific models that have high predictive value only in a particular setting (e.g., creatine kinase score for myocardial infarction), and (2) the weights assigned to specific clinical variables (i.e., their predictive value for death) vary depending on the specific disease context. For example, the effect of high blood pressure may be harmful in one condition but not in another. We found numer-

ous examples of this phenomenon during the Prospective Payment Study (Keeler et al., 1990).

The finding of superior performance of disease-specific models reinforces an earlier suggestion (Kahn et al., 1988) that hospital quality-assessment activities focus on congestive heart failure, acute myocardial infarction, stroke, and pneumonia, which collectively account for about 30 percent of in-hospital Medicare deaths (Daley et al., 1988). Moreover, these conditions are associated with significant 30-day mortality rates (about 15–25 percent), which permits clinically and statistically significant differences in outcome levels to be observed reliably (Luft and Hunt, 1986). Also, developing criteria and abstraction forms and keeping them up to date is expensive and time-consuming; scarce money and time are best spent "where the action is."

Second, *models developed using clinical data from the medical record are generally more accurate than are models that rely on administrative data* (Green et al., 1990; Iezzoni et al., 1992; Smith et al., 1991). (However, Alemi et al., 1990, observed comparable performance between clinical- and administrative-data-based models.) Whether the additional accuracy of clinical data is worth the additional costs and effort of data collection is a value judgment and will often depend on the purpose to which the results will be put.

Third, *both administrative-level models* (Smith et al., 1991; Green et al., 1991) *and clinical-level models* (Blumberg, 1991) *can be seriously biased if they are misspecified as a result of the omission of important variables, such as age.* Detection and confirmation of such biases is difficult, however, and requires knowledge both of the omitted variables and of how hospitals differ in the proportion of patients with those variables.

Finally, *only about half of the models we reviewed were tested to ensure against overfitting*. In particular, most disease-specific models were simply fit to a single sample of patients without cross-validation in a separate sample (Rodrigues et al., 1991; Fullerton et al., 1988; Bonita et al., 1988; Howard et al., 1986; Howard et al., 1989; British Thoracic Society Research Committee et al., 1987; Ortqvist et al., 1990; Celis et al., 1988; Starczewski et al., 1988; Rouleau et al., 1990; Cleland et al., 1987; Cohn and Rector, 1988; Parameshwar et al., 1992; Barin et al., 1988; Sahasakul et al., 1990; Marik et al., 1990; Fioretti et al., 1985; Waters et al., 1985; Greenland et al., 1991; Tibbits et al., 1987; Cleempoel et al., 1986). Only a few disease-specific models (Keeler et al., 1990; Daley et al., 1988; Durocher et al., 1988; Zweig et al., 1990; Pierard et al., 1989; Dubois et al., 1988; Cleempoel et al., 1988; Selker et al., 1991b) were cross-validated in separate samples of patients.

This consideration is important because statistically derived prediction models are almost always overspecific to the particular assortment of variable values contained in the original (i.e., generating) sample. This problem is more severe when many variables are included in the model. For example, Iezzoni et al. (1992) reported that models with ten or twelve variables substantially outperformed models with between 40–65 variables in separate test samples of patients. At the other extreme, Spiegelhalter (1986) reported substantially poorer cross-validation performance in a three-variable model compared to a thirteen-variable model, and Hadorn et al. (1992) observed essentially equivalent cross-validation performance between four- and eight-variable models.

These findings support Iezzoni et al.'s (1992) recent suggestion that the best mortality prediction models, in terms both of accuracy and cross-validation performance, should be disease-specific mortality prediction models consisting of a small, common core of universally important clinical variables (e.g., blood urea nitrogen, blood pressure, presences of coma), supplemented by a few disease-specific variables. The proper number of variables in these models is probably best determined empirically using cross-validation exercises, as just described.

3. Empirical Comparison of Measures

As outlined above, we initially attempted to assess and compare the predictive performance of selected mortality prediction models. So far we have seen that developers of mortality prediction models use a wide range of statistics to report model performance, thus effectively precluding performance comparisons. Clearly, therefore, one necessary step toward being able to compare model performance is for developers to report performance using a core set of "standard" statistics. In Section 4 we address the question of what those statistics should be.

There is a less tractable problem with comparing the performance of different models, however, namely, the problem of intrinsic sample predictibility. In general, samples containing patients who are relatively homogeneous (e.g., almost all quite sick or quite healthy) are more difficult to predict than are samples with patients who vary substantially in their severity of illness. Even a poor model may perform well when applied to an easy-to-predict patient population, whereas a good model may perform less well when applied to a hard-to-predict sample.

The problem of intrinsic sample predictibility is critical to the task of assessing and comparing prediction model performance. For this reason, we decided to perform an empirical evaluation of the extent to which varying the underlying distribution of explanatory variables affects the ability of different performance statistics to distinguish between a "good" model and a "poor" one. We also evaluated the effect of varying mortality rate on performance statistics.

Data for this analysis were obtained from 2,853 patients with myocardial infarction admitted at 297 randomly selected hospitals during the years 1981–82 and 1985–86.

As part of the process of developing statistical models for the RAND PPS study (Keeler et al., 1990), the total population of patients was divided into random two-thirds and one-third subsamples. The models used in evaluating the various performance measures were developed using this two-thirds "generating" set. Performance was measured both on this set and on the one-third "test" set.

We employed a variable selection process described elsewhere (Hadorn et al., 1992). A logistic regression model was developed with eight predictor variables. These variables were age, cardiac function (Killip), location of MI, acute physiol-

ogy score, systolic blood pressure score, blood SGOT level, creatine kinase level, and history of diabetes. We refer to this model as the "good model." To compare the behavior of the various performance measures on this model with their respective performance in a "poor model," we derived a complementary set of performance measures using a model that consisted of only the four weakest predictors contained in the good model. This "poor model" was only *relatively* poor; it was still statistically significant at the < 0.0005 level, with a chi-squared statistic about half that of the good model. We ran parallel analyses using both models to determine whether two or more apparently good (but in reality quite different) models could be distinguished across different data sets.

The dependent variable in our analyses was a 0/1 variable which indicated whether patients were alive or dead 30 days after admission.

Coefficients derived in the generating sets were applied to the test sets in a cross-validation exercise.

Varying Sample Distribution

After eliminating cases for which the variable representing 30-day survival was missing, the generating and test sets contained 1,728 and 860 patients, respectively. Logistic regression was performed on the generating set using the good model. Probabilities of 30-day survival were calculated for each patient. The generating and test sets were then apportioned into the following subsets:

1. Random half of patients;
2. The half of patients with below-median prediction of death;
3. The half of patients with above-median prediction of death;
4. The half of patients in the middle two quartiles of illness;
5. The half of patients in the least- and most-ill quartiles.

Performance measures were then calculated for each subset of patients using both the good and poor models. These measures included several versions of R^2,[1] c-index, percentage total correct predictions, mean predicted probability of death in patients who died (f_1) and in those who lived (f_0), and the chi-squared

[1](1) A conservative estimate based on the multiple cross-validation technique used by Daley et al. (1988), in which coefficients are recalculated for each 90 percent of the data and applied to the remaining 10 percent, (2) a Brier R^2: [(outcome index variance – Brier score)/outcome index variance], (3) a pseudo-R^2 based on the log likelihood: [LL(constant only) – LL(model)/LL(constant only)], and (4) difference by death R^2: f1 – f0.

statistic. In addition, two well-known measures of model fit were also calculated in each sample: the Hosmer-Lemeshow statistic (Lemeshow and Hosmer, 1982; Hosmer et al., 1991; Harrell and Lee., 1990) and the Brier Score and its decomposition parameters (Brier, 1950; Murphy and Winkler, 1984; and Yates, 1981).

Data were analyzed using the Stata statistical analysis program.[2] Brier Score parameters based on deciles were obtained by ranking mortality predictions.[3]

Results

Table 3 depicts the results of our analysis. Several observations are in order, beginning with a consideration of the familiar parameters R^2 and c-index. First, it can be seen that the value of both these performance measures is substantially dependent on the underlying distribution of the outcome variable (i.e., 30-day mortality). Conservative R^2, for example, ranged from 0.000 to 0.328 for the good model in the generating set. R^2 increases with the probability of death in the sample and with the spread of severity. An indication of how "easy" or "difficult" a data set is to predict can be obtained by observing this measure (or, for that matter, by observing the pattern of any other performance measure). Thus, the easiest data set is #5 (containing the extremes of illness) and the most difficult is #2 (the least ill half). Significantly, R^2 for the poor model on easy data often equals or exceeds R^2 for the good model on difficult data. The remaining measures manifest similar confounding between model performance and intrinsic sample predictibility. This finding illustrates the problem with relying on R^2 or any other statistic for assessing and comparing model performance.

When Generating and Test Sets Differ in Intrinsic Predictibility

Table 4 shows the results of a separate analysis we performed to determine the effect of developing models in a sample with substantially different underlying distributions of independent variables than the test sample.

Three distributions were examined: a set consisting of the highest and lowest quartiles of sickness (the easiest sample to predict), a random half of the sample, and the sickest half of the distribution (the most difficult of the three samples to predict). Looking down the major diagonal shows the value (shaded) of pseudo-

---

[2]Stata Reference Manual, Release 3.0. Computing Resources Center, Santa Monica, California.

[3]Stata Technical Bulletin, 11/92, pp. 20-21.

Table 3

Performance Parameters of Logistic Regression Model

		Brier Score Decomposition Parameters													
	d[a]	Outcome Index Variance d(1–d)	Brier[b] Score	Prediction Variance S	Murphy Resolution	Brier R^2	Conservative R^2	f_1	f_0	R^2 f_1–f_0	R^2 Log Likelihood	Chi Square	ROC Area	Total Correct	Hosmer-Lemeshow Statistic
Random half															
GG	.26	.19	.15	.03	.04	.24	.21	.41	.20	.21	.19	190	.76	.80	13
PG			.17	.02	.02	.11	.09	.33	.23	.10	.08	81	.70	.77	15
GT	.24	.18	.15	.03	.03	.18	.17	.61	.21	.20			.75	.79	8.9
PT			.17	.02	.02	.06	.05	.32	.24	.08			.66	.78	5.8
Least ill half															
GG	.10	.09	.09	.00	.00	.01	.00	.11	.10	.01	.02	8.2	.60	.90	9.8
PG			.09	.00	.00	.00	.00	.10	.10	.00	.01	3.7	.60	.88	9.8
GT	.12	.11	.11	.00	.00	.00	.00	.11	.10	.01			.57	.88	6.7
PT			.11	.00	.00	.00	.00	.10	.10	.00			.57	.88	9.6
Middle ill half															
GG	.18	.15	.14	.00	.01	.03	.01	.20	.17	.03	.03	27	.63	.82	6.4
PG			.14	.00	.00	.01	.00	.18	.18	.01	.01	5.6	.59	.82	6.0
GT	.21	.17	.16	.01	.01	.02	.01	.22	.19	.03			.61	.79	6.7
PT			.17	.00	.00	.00	.01	.18	.18	.00			.54	.79	4.6
Most ill half															
GG	.38	.23	.19	.06	.05	.20	.18	.50	.30	.20	.16	180	.74	.73	10
PG			.21	.02	.02	.09	.08	.43	.34	.09	.01	76	.68	.67	8.6
GT	.39	.24	.21	.04	.04	.13	.11	.47	.32	.15			.69	.70	8.4
PT			.23	.02	.03	.05	.03	.42	.35	.06			.64	.63	19
Extreme quadrants															
GG	.30	.20	.13	.05	.07	.35	.34	.54	.19	.35	.30	310	.85	.81	11
PG			.16	.04	.05	.22	.20	.45	.23	.22	.18	187	.80	.77	5.6
GT	.30	.21	.15	.05	.06	.28	.24	.49	.20	.29			.82	.79	10
PT			.17	.04	.04	.16	.15	.44	.26	.19			.76	.73	7.2

GG = good model, generating set; PG = poor model, generating set; GT = good model, test set; PT = poor model, test set.

[a] d = mortality rate.

[b] Lower scores correspond to better model performance.

Table 4

R^2 and C-Statistics When Distribution of Illness Varies from Generating to Test Set

Generating Set		Test Set: Most Ill	Test Set: Random	Test Set: Extreme
Most Ill	pseudo R^2	.16	.17	.29
	c-index	.74	.76	.84
Random	pseudo R^2	.14	.19	.29
	c-index	.73	.77	.84
Extreme	pseudo R^2	.15	.18	.30
	c-index	.74	.76	.85

NOTE: Coefficients for the "good" eight-variable model were derived from the extreme-quadrant, most ill, and random samples in turn ("generating sets") and then tested in the other two samples (test sets). Cross-validation performance varied depending on the difference in intrinsic predictability between generating and test samples.

R^2 and c-index statistics that occurs when the equation is tested on the same data set it was derived on. The diagonal value is the highest in each column, but performance with different test sets was determined almost entirely by the distribution of the test set, not by the intrinsic "goodness" of the model. Again, we conclude that internally derived statistics from a single study (or set of studies) are difficult to interpret without some indication of the underlying distribution of independent variable values (i.e., intrinsic sample predictability).

Examination of Table 4 reveals another interesting finding. Within a given test set column there is remarkable stability in the performance statistics, irrespective of which data were used to generate the model. The R^2 and c-index varied by only $\pm$ 0.01 unit within a given data subset. Thus, while measured model performance is strongly affected by the "easiness" of the test data set, our ability to produce a good model is apparently unaffected by which data are used for model development. (All of these data sets include many deaths.)

Noncomparability Example

From this analysis, it is evident that internally derived statistics cannot reliably distinguish between a good and a poor model on different data sets. The following example shows this forcefully, and shows the particular importance of sample selection. Indeed, one can show that through purposive sampling, it is always possible to make a poorer model look better than a good one.

Example. It is possible to match the performance of any severity measure S on any data set with any other severity measure T, no matter what their relative quality, on a selected subsample of patients, provided that for each value of S the latter study has patients who live and patients who die with values of T matching that value. For example, even if someone developed an excellent severity measure on unselected AMI patients, we could exactly match its performance with a severity measure based only on some irrelevant characteristic, such as shoe size, by sampling from a large enough sample of AMI patients. (This would not occur with random sampling, of course, unless the sample was sufficiently unusual.)

Proof. Let (S,D) represent a data set of severity scores and subsequent outcomes consisting of ordered pairs (S_i,D_i). Suppose that there is another population giving rise to pairs (T_j,D_j). For each observation i, sample (T_j,D_j) so that $T_j = S_i$ and $D_j = D_i$.

Since the distribution of (T_j,D_j) exactly matches the original (S_i,D_i), any function of them will also be matched.

Corollary. By purposive sampling, we can create a data set in which any severity measure discriminates perfectly. For example, we would take only people whose shoe size was greater than seven who died, and people whose shoe size was less than seven who lived. Using shoe size as a measure of severity would have $R^2 = 1$, c-index = 1, etc.

Standardization of Model Performance

To overcome the inherent noncomparability of internally derived statistics, it will be necessary to develop and apply a common, external yardstick of performance. There are at least two possible ways to do this: (1) Use one or more common data sets to which all models would be applied, and (2) compare a new model's performance against that of a standard model on the same data set.

With regard to the first possibility, *a handful of standard data sets could be identified, representing different levels of severity of illness, in which different models could be tested.* This use of common, standardized data sets is analogous to the use of biochemical standards against which clinical laboratories calibrate and test their instruments. Performance statistics would be reported for each prediction model for each data set, permitting direct comparisons of performance across different models. Just this sort of comparison was recently completed by MacKenzie and colleagues at Queens University (published results pending); Iezzoni and colleagues have recently begun a similar study at Beth Israel Hospital in Boston. Such large-scale efforts are expensive and time-consuming, however. Moreover,

the use of a common data set to standardize model performance would be difficult or impossible when the models to be compared have been developed for use in different populations. For example, some populations (or samples) might include the variables contained in the common data set; others might not.

A second, simpler standardization approach would be *to institute the use of a standard prediction model in side-by-side comparisons of model performance.* Performance reports would display the results of the model under study against that of the standard model. The use of such a standard model would also permit the direct cross-comparisons of different models using a common yardstick. (Again, however, the usefulness of this approach would be limited by the extent of commonality of variables across the standard and to-be-compared models.) A simplified version of the demographic and physiological variables collected by APACHE II[4] (e.g., a pseudo-APACHE score) would seem to be a logical choice for the standard system. Parallel evaluations using APACHE II were performed during assessment of new measures by Daley et al. (1988) and Keeler et al. (1990).

A third, weaker route to standardization is possible, in which sampling rules would be specified to minimize the effects of nonrandom sampling — and to eliminate purposive sampling. For example, sampling rules could specify that models be developed using consecutive sampling in general hospitals within specified disease or condition categories. This method, while a distinct improvement over existing practices, could not control for unsuspected differences among hospitals, however (e.g., if one hospital systematically admitted sicker patients, or patients with an unmeasured comorbidity).

As discussed above, use of a generic model will usually result in less accurate prediction for a given disease or condition than will use of a disease-specific model. Similarly, the sort of generic, standard model just described will usually provide less accurate predictions than a disease-specific measure developed on the same data. Nevertheless, a standard pseudo-APACHE-type measure should provide a reasonable approximation of the intrinsic predictibility of most patient samples, given that the variables contained in such a measure (e.g., blood pressure, renal function) represent the final common pathways of physiological instability and organ system failure.

---

[4]APACHE, which stands for Acute Physiology and Chronic Health Evaluation, is a mortality prediction model developed for use in patients admitted to medical intensive care units. The model contains several common variables related (1) to physiological stability, such as blood pressure and hemoglobin, and (2) to background mortality risk, including age. See Knaus et al. (1991).

Conclusions

On the basis of the foregoing considerations and analyses, we make the following conclusions and recommendations:

1. Mortality prediction models should be disease- or condition-specific. Models should generally consist of a core of common clinical and demographic variables (e.g., age, blood pressure), supplemented by a few disease-specific variables.
2. Most existing models are sufficiently accurate to adjust for differences in case mix *provided that an adequate number of cases are sampled* from each provider (e.g., hospital or physician). A few hundred cases per provider would be a reasonable minimum given commonly reported outcome (i.e., mortality) rates.
3. Because of pervasive differences with respect to the methods used to test and report model performance, we are unable to draw conclusions concerning the relative performance of existing models. Standardization of performance statistics would be highly desirable (see Section 2).
4. Even if performance statistics are standardized, however, model comparisons would be continue to be hampered by differences in the intrinsic predictability of the samples used to develop and test models. For this reason, one or more external, common prediction models should be developed to control for sample predictability and to permit the meaningful assessment and comparison of model performance.

4. Expert Meeting

To obtain peer review of the work described in Section 1, and to extend our preliminary conclusions, an earlier version of that material was mailed to about 30 experts in health outcome prediction. A subset of 14 scholars (listed in Appendix A) attended a meeting at RAND in Washington, D.C., hosted by RAND and HCFA staff, to discuss the possible recommendation of guidelines for the development, testing, and reporting of mortality prediction models. The agenda included the question of appropriate statistical performance measures and intrinsic sample predictability discussed in Sections 1–3 of this report. In addition, the agenda addressed data-collection issues, including timing and variable selection, and the feasibility and usefulness of models, including problems resulting from providers attempting to "game" the system. To focus the debate, the discussion was restricted to the use of mortality prediction models for adjusting hospital death rates in quality assessment efforts.

Initial Recommendations

Panelists were initially asked to suggest items or criteria that they believed should be included in all new studies of mortality prediction models. These criteria were to be considered analogous to the specification of inclusion and exclusion criteria in clinical studies.

A total of 33 items were suggested for inclusion as issues to be addressed during the development, testing, and reporting of mortality prediction models. These items, which are listed in Appendix B, fell into three broad groups: performance statistics, data-collection issues, and model feasibility and usefulness.

The panel was then divided into two subgroups. One discussed data-collection issues; the other, issues related to performance statistics. Lack of time precluded a separate discussion of model feasibility and usefulness, although several of these issues were discussed in the context of data collection (see below). After conducting their separate discussions, representatives of the two groups presented their conclusions to the full panel. These conclusions constitute recommendations that are intended to apply only to major or significant model development activities, such as severity adjustment systems designed to be used for such purposes as adjusting hospital death rates.

Recommendations of the Data-Collection Subgroup

The data-collection subgroup was charged with considering such issues as timing of data collection, variable selection, appropriate data sources, and protocols for handling missing data. The format was unstructured and participants were free to consider whatever issues they deemed most important for purposes of advancing the science of mortality prediction.

The subgroup's recommendations were as follows:

(a) The outcome of interest should be death within 30 or 90 days post-admission (or post-surgery) unless there is good justification for using in-hospital death. The problem with using in-hospital death rates to measure performance is that hospital policy can and does change the location of deaths. In-hospital deaths will be higher for hospitals that do not discharge patients to die, even if the patient outcomes are good on average (Jencks et al., 1988). Some panelists recommended that only if 90–95 percent of all 30-day deaths occur in hospital should in-hospital death be considered a sufficient endpoint. Other panelists were uncomfortable with attempting to specify a given cutoff point for making this decision. Clearly, however, the benefit of using 30-day mortality may not justify substantial additional costs for data collection if a high proportion of deaths occur in hospital. On the other hand, hospital mortality is an insufficient endpoint if patients often die within a short period of time posthospital.

Reporting time to death (in days) was considered preferable by many panelists, provided that fixed-time (e.g., 30-day) death rates are also reported or enough information is provided to permit calculation of these rates. Other panelists expressed the concern that the use of time to death as the dependent variable (e.g., using Cox regression) could produce misleading and peculiar results. The distinction between, for example, a death on day 1 and a death on day 29 is likely to be of little relevance from the perspective of quality assessment. A hospital that postpones an inevitable death by three or four weeks (whether or not this was done in an attempt to improve its outcome statistics) should not be considered "better" than a hospital that allows a terminal case to die early. This is one of several examples identified during the meeting of how perverse provider incentives might be created if great care is not taken in how expected outcomes are calculated.

(b) Mortality must be a meaningful measure of outcome for the condition in question. Diseases or conditions with < 2 percent mortality rate (e.g., treatment of back pain and cataract surgery) should not be studied using mortality as an outcome. In these populations, death could be considered a "sentinel event,"

that is, any occurrence could trigger an investigation (Rutstein et al., 1976). At the other extreme, patients admitted for terminal care (i.e., who are expected to die) should not be included in mortality prediction analyses.

(c) The relevance of selected variables should be supported by the literature. Where available, existing literature reviews should provide the primary basis for model selection. When necessary, new literature reviews could be performed to supplement clinical experience for determining relevant variables. Variables should be precisely defined, including (where possible and relevant) how they differ from others' models. Model developers should also cite reasons for omitting any apparently relevant predictors variables, e.g., the variable is too difficult or expensive to collect or previous studies have shown that the variable is unnecessary or undesirable. Examples of the latter problem might include problems of coding reliability (e.g., poor reliability of coding "urgent" vs. "emergent" admissions) or confounding with outcomes (e.g., use of "cardiac arrest" as predictor variable for death).

(d) There should be clinical face validity for any variable included in the model, including a clinical theory of why it is important. Panelists encouraged a priori hypotheses but would accept post-hoc theories consistent with findings. An attempt should be made to explain counterintuitive variables clinically. In addition, other potential explanations should be sought for observed counterintuitive relationships, e.g., selection bias. Note that this recommendation applies only to the development of final risk-adjustment systems, not to the development and testing of new models by researchers. If researchers could not test variables that lacked evidence of predictive validity, the state of knowledge would never advance. Consistent correlation of an unsuspected variable with death could then lead to new theories and experiments attempting to explain this association. Pending completion of such studies, however, it is probably best to restrict the variables in most mortality prediction models to those with evident face validity.

(e) Independent variables should occur frequently enough to be meaningful. The appropriate frequency will vary according to that variable's strength as a predictor. For example, a comorbidity that is present in 2 percent of deaths but only 0.001 percent of survivors might be appropriate for inclusion in the model, whereas a comorbidity that is present in 2 percent of deaths and 1 percent of survivors would not. The cost of collecting an item does not depend on prevalence, but the benefit of collecting the variable may.

(f) Both administrative data and medical records data (including hard-copy or electronic data) are appropriate data sources. In either case, data used for prediction should be drawn from among those routinely collected for patient care.

(Again, this recommendation applies to final risk-adjustment systems used for such purposes as those listed in Table 1. In the research context, there is nothing wrong with using data that are not routinely collected.) Variables included in the model should be routinely required for clinical decisionmaking and should be justified as being helpful to patient care. Otherwise, hospitals may perform clinically unnecessary tests (to make their patients look sicker) and hospitals that do not do so may show worse risk-adjusted outcomes (because the acuity of their patients' conditions is not completely captured).

(g) Missing data elements should be coded as normal (during data analysis, not during construction of the data base itself) unless there are a priori reasons or well-described methods for imputing other values for such data elements. For example, it might be reasonable to impute a value for diastolic blood pressure if a hospital reports only the systolic pressure, or an abnormal blood gas value could be imputed in a patient with low oxygen saturation by pulse oximetry.

Some reasons for missing data include: (1) The patient is admitted for terminal, comfort care only and the expected tests are not done (as noted above, these patients should be excluded from mortality analyses); (2) some tests are done only if patients look sick, in which case a missing value might be indicative of less-ill patients (tested in Keeler et al., 1990); and (3) the patient may have died before the data could be collected.

A strategy of coding missing data as normal will generally penalize hospitals to the extent they fail to record data that reflect severity of illness, because patients will not look as sick as they really are and death will appear to be a relatively low probability. This assumption may not always apply in practice, however, especially if coding accuracy/thoroughness or test frequency are positively correlated with severity. Sensitivity analysis can be performed to determine the effect of coding missing data as normal.

(h) All data should be collected in the first two calendar days, i.e., within 48 hours (up to two weeks of preadmission preoperative laboratory data can be included). In addition, prehospital variables may be important and appropriate to include in the model, such as "admitted from nursing home" or "functional status." The rationale for the 48-hour window (rather than a 24-hour window, which would be more desirable) is that data-collection systems often list only the date of an occurrence. Therefore, the first 24-hour period falls almost entirely on hospital day #1 for a patient admitted just after midnight, whereas it falls mostly on hospital day #2 for a patient admitted in the evening. The initial diagnostic work-up is still under way at midnight on hospital day #1 for the latter patient.

Thus, a time window that captures 24 hours of care for every patient inevitably captures up to 48 hours of care for a subset.

The panelists encouraged the earliest possible data collection. First available data should be used, unless there is a medically justifiable reason for another choice, e.g., an initially normal temperature was due to a recent dose of Tylenol. To accommodate such cases, some panelists advocated a strategy of using the worst value observed within first 24 hours, noting that in the great majority of cases, the worst value will also be the first. It is possible, however, that a strategy of "first or worst" might require more training of chart abstractors and could be less reliable than a "first only" strategy.

(i) Mortality should not be an independent variable. Proxies for mortality, e.g., asystole, apnea, pulselessness, cardiac arrest, or no blood pressure, should be used only if they exist prehospital. As a corollary, since mortality is the dependent variable, it is inappropriate to exclude a subset of deaths, such as deaths on the day of admission, from analysis.

(j) The DNR (do not resuscitate) variable is a legitimate predictor if coded within the first 48 hours. Concern was expressed about bias or gaming; for example, providers could (consciously or unconsciously) tend to write DNR orders when death was expected, to make patients look sicker and thus not be "dinged" for an unexpected death. More research is encouraged in this area. Indeed, some panelists felt strongly that DNR orders should not be recommended for inclusion in mortality prediction models until additional research has been conducted to determine the likely effect of such inclusion on hospital and physician behavior. As noted above, patients admitted only for terminal care or comfort care should be excluded from the sample.

(k) The precision and reproducibility of measurement of data elements (e.g., are blood pressure monitors uniformly accurate) should be commented on in discussion of measurement and variables.

(l) Objectivity or subjectivity of variable collection should be addressed (e.g., degree to which data are subject to bias or manipulation). For variables that are subject to these problems, studies of their expected relationships to other known objective variables should be performed. Also, it is desirable to study whether data coding is consistent across hospitals. Is it better in hospitals that have better outcomes? Measures of reliability (e.g., kappa, R_{ii}) should be calculated for each item, and the sample should be large enough to make this statistic valid.

(m) Studies should report the characteristics of hospitals (e.g., size, location, amount of teaching) that supplied the data to generate a prediction model. In-

clusion and exclusion criteria should be specified. Performance statistics should be reported in a separate sample of patients from the one upon which the model was based. (The major concern here is that developers of risk-adjustment systems may apply or market their systems without demonstrating external validity. A system developed on teaching hospital patients in the Northeast may not work well with community hospital patients in the Midwest, and so forth.)

(n) At the time of bidding/buy decisions, and at any time during use of the system, potential users should have all relevant system information made available, including coefficients, weights, and equations. These should be verifiable by the user (or by experts selected by the user). Where models are developed with private money, all of this information should be respected as proprietary.

Additional ideas concerning data collection were presented during open discussion:

(o) Inter-rater reliability of the predictions themselves should be reported (not just the reliability of data coding).

(p) Model developers need to pay attention to diagnostics. Are there a few outliers that affect everything? Statements such as "We checked for overly influential observations and made sure that the scaling of the variables was appropriate" are comforting to the reader/user, assuming that an explicit description is provided of how this was done.

Recommendations of the Performance Statistics Subgroup

Recommendations emerging from the performance statistics subgroup were designed to apply to all substantial efforts to develop new models. The group was charged with recommending specific statistics for assessing and comparing model performance, as well as formulating recommendations concerning the standardization of model performance and assessment of cross-validation performance.

The performance statistics subgroup recommended that the following statistics be routinely reported:

(a) C-index (area under the ROC curve). This is the universal choice for a measure of discrimination, i.e., the ability of the model to distinguish between patients who will survive and those who will die.

(b) In addition to discrimination, models should be well calibrated. For example, roughly 10 percent of patients assigned a probability of death of 0.1 should die, and the great majority of patients assigned probabilities of death of 0.8–0.9 should die. For measuring calibration, studies should report the number of expected vs. observed deaths stratified by deciles. Deciles should be calculated both according to risk (with a tenth of the sample in each risk category) and by fixed cutoff points (e.g., probability of death = 0–0.1, 0.1–0.2, etc.). The Hosmer-Lemeshow statistic and decompositions of the Brier Score (see p. 16) are examples of appropriate approaches to this calculation. Publication of a calibration curve is also recommended (see Figure 1 for an example (from Selker et al., 1991a), ideally with a histogram underneath the curve to depict the distribution of the sample.

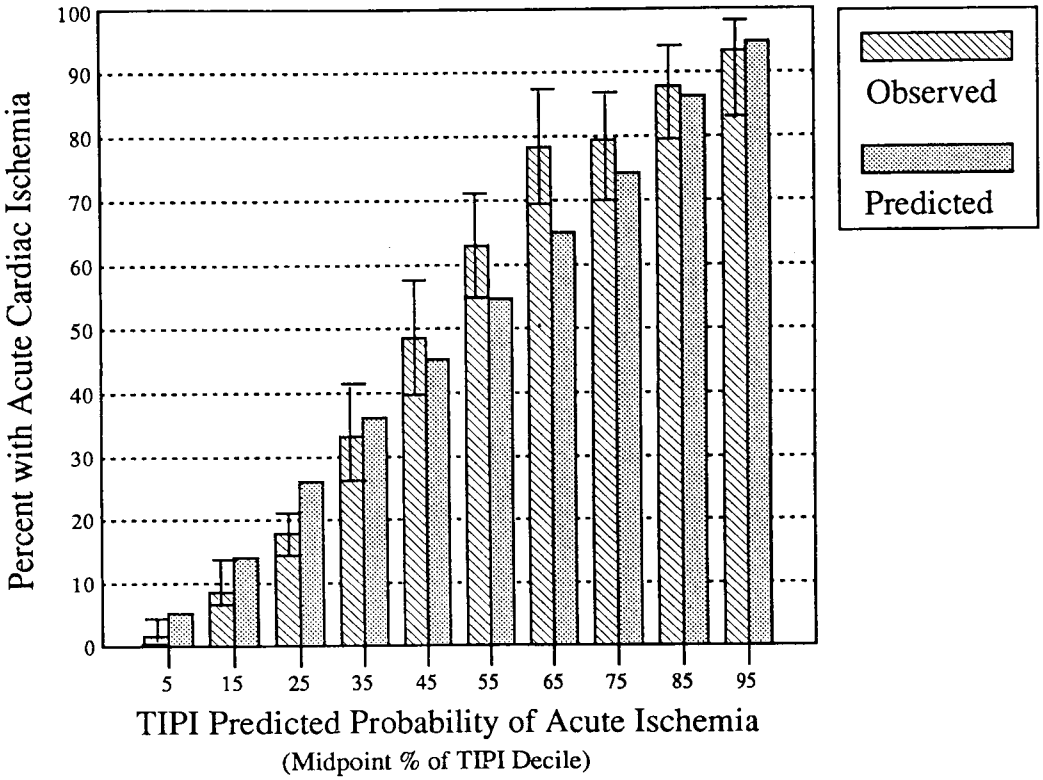

Figure 1—Example of a calibration curve based on deciles. From Selker et al. (1991a), p. 615. ©J. B Lippincott Company. Reprinted with permission.

There was considerable controversy concerning the usefulness and appropriateness of R^2 as a summary performance statistic. However, most panelists ap-

peared to believe that R^2 provides useful summary information and should be included with the other statistics just mentioned.

(c) Models should be tested alongside a common, external model consisting of a core set of variables including, probably, age, sex, physiological/laboratory data, and one or more disease-specific elements. This practice will provide a common yardstick with which to adjust for ease or difficulty of predicting outcomes in given data sets. (See the discussion and analysis in Section 1 of this report.) Much work needs to be done before this recommendation can be implemented. Some questions emerging from the panel were: Who would develop the proposed common, external model? Would it differ for different types of patients (e.g., surgical vs. medical vs. pediatric)? What if all of the basic variables are not available in a data set used to develop a proposed system (e.g., administrative data without physiological variables)? How would the comparison of models be assessed (e.g., likelihood ratio test, R^2, c-index)?

(d) Some statement should always be included about how the investigators protected against overfitting (e.g., boot-strapping, split-sample, number of variables ever tried vs. number of least frequent outcome, cross-validation in separate populations). Any of these approaches is acceptable.

The panel agreed that the recommendations developed during this meeting should be aimed primarily at: (1) the research community, (2) funders of research in this area, (3) system users, and (4) journal editors. However, we believe that the promulgation and acceptance of these recommendations is also important for the thousands of physicians and hospitals whose quality of care may be judged according to mortality prediction models, and to millions of patients whose access to care and selection of treatment may increasingly depend on statistical predictions of outcome.

5. Principal Conclusions and Recommendations

The principal conclusions and recommendations resulting from this project are as follows:

1. Despite years of work, the science of mortality prediction is still in its early stages of development. Judging by observed performance statistics, however, the accuracy (or validity) of most existing mortality prediction models appears adequate for group-level assessments, *provided* that (1) sufficient numbers of patients are sampled from each provider and (2) sampling is performed in a reasonably representative manner, preferably randomly. Conclusions regarding quality of care based on statistics alone should always be drawn with extreme caution, and should not be drawn at all unless at least a few hundred cases per provider are included in the sample (see the discussion on pp. 10–11). In most cases, discrepancies between observed and expected outcomes should serve to trigger a review of the process of care to determine if quality problems in fact exist.

2. Comparative performance of prediction models cannot be reliably assessed at the present time, in part because a wide variety of different performance statistics are currently being used. For this reason, all studies should report the c-index and some measure of calibration, such as the Hosmer-Lemeshow statistic and Brier Score parameters referred to on p. 16.

3. An even more fundamental problem with model comparison is that model performance strongly depends on the intrinsic "predictability" of patient samples. For this reason, one or more external, common models should be developed for use in estimating intrinsic sample predictability. Use of such standardized models would greatly enhance our ability to compare model performance. Future research should address the development, testing, and use of standardized prediction models.

4. In the meantime, we recommend that all models be tested alongside a pseudo-APACHE-type model containing several common physiological and laboratory variables known to be related to the risk of mortality. Statistics from this model should be reported along with those of the actual model of interest.

5. Mortality prediction models are almost always overspecific for the patient samples upon which they were developed, and thus performance usually deteri-

orates when models are applied to different patient samples (depending on the underlying distribution of the sample—see Tables 3 and 4). For this reason, we recommend that mortality prediction models always be tested in patient samples distinct from those in which the models were developed. Performance statistics should be reported from this separate test sample.

6. A wide variety of sampling and data-collection issues are of critical importance in ensuring the reliability and validity of mortality prediction models in assessing quality of care. Among these issues are the timing of data collection, variable selection, and treatment of missing data.

7. All substantial research or commercial efforts to develop mortality prediction models should endeavor to incorporate the recommendations developed during this project, as summarized in Section 4.

Appendix A
List of Participants at Expert Meeting

Arlene Ash, Ph.D.
Health Care Research Unit
Boston, MA

Mark S. Blumberg, M.D.
Director of Special Studies
Kaiser Foundation Health
Plan, Inc.
Oakland, CA

Neal V. Dawson, M.D.
MetroHealth Medical Center
Baltimore, MD

Frank Harrell, Ph.D.
Associate Professor of Biostatistics
Duke University
Durham, NC

Robert Hauchens, Ph.D.
Senior Statistician
Systemetrics
Santa Barbara, CA

Susan Horn, Ph.D.
Intermountain Healthcare
Salt Lake City, UT

Lisa Iezzoni, M.D.
Division of General Medicine
Beth Israel Hospital
Boston, MA

Stephen F. Jencks, M.D., MPH
HCFA - Office of Research
Baltimore, MD

Stanley Lemeshow, Ph.D.
Chair of Biostatistics Department
School of Public Health
University of Massachusetts

Mark Moskowitz, M.D.
Chief of General Internal Medicine
University Hospital
Boston, MA

Patrick Romano, M.D., MPH
Assistant Professor of Medicine
Division of General Medicine
University of California, Davis
Sacramento, CA

Harry P. Selker, M.D., MSPH
Multicenter Cardiology & Health
Services Research Unit
New England Medical Center
Boston, MA

J. William Thomas, Ph.D.
Associate Professor
School of Public Health
University of Michigan
Ann Arbor, MI

Douglas P. Wagner, Ph.D.
Senior Research Scientist
ICU Research Unit
George Washington Medical Center
Washington, DC

Project Officer
Harry Savitt, Ph.D.
Health Care Financing Administration
Baltimore, MD

RAND
Robert Brook
David Hadorn
Emmett Keeler
William Rogers

Consultants are listed for information purposes only. None of these consultants necessarily agrees with any particular conclusion or recommendation contained in this document.

Appendix B
Expert Panel's Initial Recommendations for Evaluation Criteria

As noted in the text (p. 22), the expert panel meeting convened in Phase II of this project began with a nominal group process in which panelists took turns listing what they believed should be standard criteria addressed in studies that develop and test mortality prediction models. The suggested criteria fell broadly into three categories: issues pertaining to (1) performance statistics, (2) data collection and variable selection, and (3) model feasibility and usefulness. Many of the criteria are expanded upon in Section 4 of this report.

Performance Statistics

1. Model performance should be measured in an independent data set.
2. A description should be provided of how overfitting of the model was avoided (e.g., cross-validation, boot-strapping).
3. Both calibration and discrimination should be reported.
4. Summary statistics (e.g., Brier Score) should be decomposed.
5. Confidence intervals or standard errors around estimates should be provided and should be adjusted for stepwise variable selection and all other assessment of predictors (e.g., boot-strap).
6. The effect of potential biases created by flaws in data collection (e.g., missing values or variables) should be modeled.
7. Intrinsic predictability of sample should be assessed using external, common prediction model.
8. Model developers should report how much extra variance is explained by their model after breaking the sample into disease categories (which will explain much of the variance in and of itself).
9. Model should be tested using sufficient numbers of patients per hospital to permit adequate power to show hospital differences.
10. Possibility of interactions between severity and quality should be addressed.

11. Performance statistics should be compared with maximum achievable rather than some unachievable limit (e.g., $R^2 = 1$).
12. Reliability of model in actual use (e.g., inter-rater reliability) should be reported.
13. Cell sizes should be specified for relevant subgroups, i.e., how many patients had each combination of predictor variables.

Data Collection/Selection

1. Allowed predictor variables should be logical.
2. Protocol for handling missing data should be specified.
3. Variables should be easy to understand and interpret, particularly summary parameters (e.g., derived through factor analysis).
4. Development or generating sample should be broad based.
5. Assumption of homogeneity within diagnostic categories or diseases should be tested.
6. Appropriateness of selected outcome should be addressed, because death may not reflect quality in all circumstances.
7. Extension of data collection to the patient or the physician, i.e., not just from the medical record, should be considered.
8. Data-collection timing issues should be addressed, e.g., first vs. worst values, selected time of outcome, handling of the actual date of death when the data specify only who died during a given month, e.g., Social Security data.

Feasibility and Usefulness

1. All data, variables, and equations should be made explicit.
2. Standard diagnostic groups should be used.
3. Cost, time required, and personnel needed for coding should be specified.
4. How procedure-dependent is the model (e.g., data elements that are affected by what was done to/for patient)?
5. Subjectivity, verifiability, and "fudgeability" of data elements should be addressed.
6. Degree of external validity (e.g., resistance to different coding practices) should be estimated.

7. Model performance should be evaluated over time.
8. Models should be developed to predict what should happen with optimal care, not average care.
9. "Face usefulness" should be determined, i.e., does model contribute to medical care generally?
10. Currency of system should be addressed (e.g., what happens if the system used for coding data changes)?
11. The purpose of the model should be explicit.
12. Possible effect of the standards on hospital behavior should be addressed.

Bibliography

Alemi F, Rice J, and Hankins R. "Predicting In-Hospital Survival of Myocardial Infarction: A Comparative Study of Various Severity Measures." *Medical Care* 1990; 28:9; 762–775.

Barin E, Lister VJ, Jones MP, et al. "A Clinical Model for Predicting Survival Following Acute Myocardial Infarction in Patients without Cardiogenic Shock: A Multivariate (Cox) Analysis." *Australian New Zealand Journal of Medicine* 1988; 18:61–66.

Blumberg M. "Biased Estimates of Expected Acute Myocardial Infarction Mortality Using Medisgroups Admission Severity Groups." *Journal of the American Medical Association* 1991; 265:2965–2970.

Bonita R, Ford MA, and Stewart AW. "Predicting Survival After Stroke: A Three-Year Follow-Up." *Stroke* 1988; 19:6; 669–673.

Brier GW. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 1950; 75:1–3.

British Thoracic Society Research Committee. "Community-Acquired Pneumonia in Adults in British Hospitals in 1982–1983: A Survey of Aetiology, Mortality, Prognostic Factors and Outcome." *Quarterly Journal of Medicine* 1987; 62:239; 195–220.

Celis R, Torres A, Gatell JM, et al. "Nosocomial Pneumonia: A Multivariate Analysis of Risk and Prognosis." *Chest* 1988; 93:2; 318–324.

Charlson M, Pompei P, Ales K, et al. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation." *Journal of Chronic Disease* 1987; 40:5; 373–383.

Charlson M, Sax F, MacKenzie CR, et al. "Assessing Illness Severity: Does Clinical Judgment Work?" *Journal of Chronic Disease* 1986; 39:6; 439–454.

Cleempoel H, Vainsel H, Bernard R, et al. "Predictors of Early Death after Acute Myocardial Infarction: Two Months Follow-Up." *European Heart Journal* 1986; 7:305–311.

Cleempoel H, Vainsel H, Dramaix M, et al. "Limitations on the Prognostic Value of Predischarge Data after Myocardial Infarction." *British Heart Journal* 1988; 60:98–103.

Cleland JGF, Dargie HJ, and Ford I. "Mortality in Heart Failure: Clinical Variables of Prognostic Value." *British Heart Journal* 1987; 58:572–582.

Cohn JN, and Rector TS. "Prognosis of Congestive Heart Failure and Predictors of Mortality." *The American Journal of Cardiology* 1988; 62:25A–30A.

Daley J, Jencks S, Draper D, et al. "Predicting Hospital-Associated Mortality for Medicare Patients; A Method for Patients with Stroke, Pneumonia, Acute Myocardial Infarction, and Congestive Heart Failure." *Journal of the American Medical Association* 1988; 260:24; 3617–3624.

DesHarnais S, Chesney J, Wroblewski M, et al. "The Risk-Adjusted Mortality Index: A New Measure of Hospital Performance." *Medical Care* 1988; 26:12; 1129–1145.

Detre K, Peduzzi P, Murphy M, et al. "Effect of Bypass Surgery on Survival in Patients in Low- and High-Risk Subgroups Delineated by the Use of Simple Clinical Variables." *Circulation* 1981; 63(6): 1329–1338.

Dubois C, Pierard LA, Albert A, et al. "Short-Term Risk Stratification at Admission Based on Simple Clinical Data in Acute Myocardial Infarction." *American Journal of Cardiology* 1988; 61:216–219.

Durocher A, Saulnier F, Beuscart R, et al. "A Comparison of Three Severity Score Indexes in an Evaluation of Serious Bacterial Pneumonia." *Intensive Care Medicine* 1988; 14:39–43.

Fioretti P, Brower RW, Simoons ML, et al. "Prediction of Mortality During the First Year after Acute Myocardial Infarction from Clinical Variables and Stress Test at Hospital Discharge." *American Journal of Cardiology* 1985; 55:1313–1318.

Fullerton KJ, MacKenzie G, and Stout RW. "Prognostic Indices in Stroke." *Quarterly Journal of Medicine* 1988; 66:250; 147–162.

Green J, Passman LJ, and Wintfeld N. "Analyzing Hospital Mortality: The Consequences of Diversity in Patient Mix." *Journal of the American Medical Association* 1991; 265:1849–1853.

Green J, Wintfeld N, Sharkey P, et al. "The Importance of Severity of Illness in Assessing Hospital Mortality." *Journal of the American Medical Association* 1990; 263:2; 241–246.

Greenfield S, and Morgenstern H. "Ecological Bias, Confounding, and Effect Modification." *International Journal of Epidemiology* 1989; 18:269–274.

Greenland P, Reicher-Reiss H, Goldbourt U, et al. "In-Hospital and 1-Year Mortality in 1,524 Women after Myocardial Infarction: Comparison with 4,315 Men." *Circulation* 1991; 83:2; 484–491.

Hadorn DC, Draper D, Rogers WH, et al. "Cross Validation Performance of Mortality Prediction Models." *Statistics in Medicine* 1992; 11: 475–489.

Hannan EL, Kilburn H Jr., O'Donnel JF, et al. "Adult Open Heart Surgery in New York State. An Analysis of Risk Factors and Hospital Mortality Rates." *Journal of the American Medical Association* 1990; 264:2768–2774.

Harrell FE Jr., and Lee KL. "Using Logistic Model Calibration to Assess the Quality of Probability Predictions." unpublished paper, 1990.

Horn S, Sharkey P, Buckle J, et al. "The Relationship Between Severity of Illness and Hospital Length of Stay and Mortality." *Medical Care* 1991; 29:4; 305–317.

Hosmer DW, Taber S, Lemeshow S. "The Importance of Assessing the Fit of Logistic Regression Models: A Case Study." *American Journal of Public Health* 1991; 81:12; 1630–1635.

Howard G, Evans GW, Murros KE, et al. "Cause of Specific Mortality Following Cerebral Infarction." *Journal of Clinical Epidemiology* 1989; 42:1; 45–51.

Howard G, Walker MD, Becker C, et al. "Community Hospital-Based Stroke Programs: North Carolina, Oregon, and New York. III. Factors Influencing Survival after Stroke: Proportional Hazards Analysis of 4219 Patients." *Stroke* 1986; 17:2; 294–299.

Iezzoni LI, Ash AS, Coffman G, et al. "Admission and Mid-Stay Medisgroups Scores as Predictors of Death Within 30 Days of Hospital Admission." *American Journal of Public Health* 1991; 812:74–78.

Iezzoni L, Ash A, Coffman G, et al. "Predicting In-Hospital Mortality. A Comparison of Severity Measurement Approaches." *Medical Care* 1992; 30.

Jencks SF, Williams DK, and Kay TL. "Assessing Hospital-Associated Deaths from Discharge Data: The Role of Length of Stay and Comorbidities." *Journal of the American Medical Association* 1988; 260: 2240–2246.

Kahn KL, Brook RH, Draper D, et al. "Interpreting Hospital Mortality Data: How Can We Proceed?" *Journal of the American Medical Association* 1988; 260:24; 3625–3628.

Kahn KL, Rogers WH, Rubenstein LV, Sherwood MJ, Reinisch EJ, Keeler EB, Draper D, Kosecoff J, and Brook RH. "Measuring Quality of Care with Explicit Process Criteria Pre- and Post-Implementation of the DRG-Based Prospective Payment System." *Journal of the American Medical Association* 1990; 264:1969–1973.

Keeler EB, Kahn KL, Draper D, et al. "Changes in Sickness at Admission Following the Introduction of the Prospective Payment System." *Journal of the American Medical Association* 1990; 264:15; 1962–1968.

Keeler EB, Rubenstein LV, Kahn KL, et al. "Hospital Characteristics and Quality of Care." *Journal of the American Medical Association* 1992; 268:13; 1709–1714.

Knaus WA, Wagner DP, and Draper EA. "The APACHE III Prognostic System: Risk Prediction of Hospital Mortality for Critically Ill Hospitalized Adults." *Chest* 1991; 100:1619–1636.

Knaus WA, Draper EA, Wagner DP, et al. "Apache II: A Severity of Disease Classification System." *Critical Care Medicine* 1985; 13:818–824.

Lemeshow S, et al. "A Method for Predicting Survival and Mortality of ICU Patients Using Objectively Derived Weights." *Critical Care Medicine* 1985; 13:519–525.

Lemeshow S, and Hosmer DW Jr. "A Review of Goodness of Fit Statistics for Use in the Development of Logistic Regression Models." *American Journal of Epidemiology* 1982; 115:1; 92–106.

Luft HS, and Hunt SS. "Evaluating Individual Hospital Quality Through Outcome Statistics." *Journal of the American Medical Association* 1986; 255:2780–2784.

Marik PE, Lipman J, Eidelman IJ, et al. "Clinical Predictors of Early Death in Acute Myocardial Infarction: A Prospective Study of 233 Patients." *South African Medical Journal* 1990; 77:179–182.

Meyer H. "Cleveland Starts Hospital Quality Project." *American Medical News* 1990; 6–7.

"Minnesota Blues' Payment Plan Is the First to Tie Reimbursement to Outcome." *Outcomes Measurement and Management* 1991; 2:1–2.

Murphy A and Winkler R. "Probability Forecasting in Meteorology." *Journal of the American Statistical Association* 1984; 79:387; 489–500.

Nemes J. "Capital Borrowings May Require Proof Of Quality." *Modern Healthcare* 1991; 44.

Ortqvist A, Hedlund J, Grillner L, et al. "Aetiology, Outcome and Prognostic Factors in Community-Acquired Pneumonia Requiring Hospitalization. " *European Respiratory Journal* 1990; 3:1105–1113.

Parameshwar J, Keegan J, Sparrow J, et al. "Predictors of Prognosis in Severe Chronic Heart Failure." *American Heart Journal* 1992; 123:2; 421–426.

Park RE, Brook RH, Kosecoff J, et al. "Explaining Variations in Hospital Death Rates." *Journal of the American Medical Association* 1990; 264: 484–490.

Pierard LA, Cubois C, Albert A, et al. "Prediction of Mortality After Myocardial Infarction by Simple Clinical Variables Recorded During Hospitalization." *Clinical Cardiology*. 1989; 12:500–504.

Rodrigues CJ and Joshi VR. "Predicting the Immediate Outcome of Patients with Cerebrovascular Accident: A Prognostic Score." *Journal of Association of Physicians of India* 1991; 39:2; 175–180.

Rouleau J, Shenasa M, Champlain J de, et al. "Predictors of Survival and Sudden Death in Patients with Stable Severe Congestive Heart Failure due to Nonischemic Causes: A Prospective Long Term Study of 200 Patients." *Canadian Journal of Cardiology* 1990; 6:10; 453–460.

Rubenstein LV, Kahn KL, Reinish EJ, Sherwood MJ, Rogers WH, Kamberg C, Draper D, and Brook RH. "Changes in Quality of Care in the United States Between 1981 and 1986 for Five Diseases Measured by Implicit Review." *Journal of the American Medical Association* 1990; 264:1974–1979.

Rutstein DD, et al. "Measuring the Quality of Care—A Clinical Method." *New England Journal of Medicine* 1976; 294: 582–588

Sahasakul Y, Chaithiraphan S, Panchavinnin P, et al. "Multivariate Analysis in the Prediction of Hospital After Acute Myocardial Infarction." *British Heart Journal* 1990; 64:182–5.

Selker HP, Griffith JL, and D'Agostino RB. "A Tool for Judging Coronary Care Unit Admission Appropriateness, Valid for Both Real-Time and Retrospective Use: A Time-Insensitive Predictive Instrument (TIPI) for Acute Cardiac Ischemia: A Multicenter Study. "*Medical Care* 1991a; 29:7; 610–617.

Selker HP, Griffith JL, and D'Agostino RB. "A Time-Insensitive Predictive Instrument for Acute Myocardial Infarction Mortality: A Multicenter Study." *Medical Care* 1991b; 29:12; 1196–1211.

Smith DW, Pine M, Bailey RC, et al. "Using Clinical Variables to Estimate the Risk of Patient Mortality." *Medical Care* 1991; 29:1108–1129.

Spiegelhalter DJ. "Probabilistic Prediction in Patient Management and Clinical Trials." *Statistics and Medicine* 1986; 5:421–433.

Starczewski AR, Allen SC, Vargas E, et al. "Clinical Prognostic Indices of Fatality in Elderly Patients Admitted to Hospital with Acute Pneumonia." *Age and Ageing* 1988; 17:181–186.

Teres D, Lemeshow S, Avrunin J, et al. "Validation of the Mortality Prediction Model for ICU Patients." *Critical Care Medicine* 1987; 15:3; 208–212.

Thomas J and Ashcraft M. "Measuring Severity of Illness: A Comparison of Interrater Reliability Among Severity Methodologies." *Inquiry* 1989; 26; 483–492.

Tibbits PA, Evaul JE, Goldstein RE, et al. "Serial Acquisition of Data to Predict One-Year Mortality Rate After Acute Myocardial Infarction." *American Journal of Cardiology* 1987; 60:451–5.

Wasson JH, Sox HC, Neff RK, et al. "Clinical Prediction Rules. Applications and Methodological Standards." *New England Journal of Medicine* 1985; 313:13; 793–797.

Waters DD, Bosch X, Bourchard A, et al. "Comparison of Clinical Variables and Variables Derived from a Limited Predischarge Exercise Test as Predictors of Early and Late Mortality after Myocardial Infarction." *Journal of the American College of Cardiology* 1985; 5:1; 1–8.

Weingarten S, Bolus R, Riedinger M, et al. "The Principle of Parsimony: Glasgow Coma Scale Score Predicts Mortality as Well as the APACHE II Score for Stroke Patients." *Stroke* 1990; 21:9; 1280–1282.

Yates JF. "External Correspondence: Decompositions of the Mean Probability Score." *Organizational Behavior and Human Performance* 1981; 30; 132–156.

Zweig S, Lawhorne L, and Post R. "Factors Predicting Mortality in Rural Elderly Hospitalized for Pneumonia." *The Journal of Family Practice* 1990; 30:2; 153–159.