oshpd

Office of Statewide Health Planning and Development

# California Intensive Care Outcomes Project *(CALICO)*

## First Year Report

October 2004

***Arnold Schwarzenegger***
Governor
State of California

***S. Kimberly Belshé***
Secretary
Health and Human Services Agency

***David M. Carlisle, M.D., Ph.D.***
Director
Office of Statewide Health Planning and Development

**First Year Report**

# CALIFORNIA INTENSIVE CARE OUTCOMES (CALICO) PROJECT

**OCTOBER 2004**

# CALIFORNIA INTENSIVE CARE OUTCOMES (CALICO) PROJECT


## Office of Statewide Health Planning and Development

Joseph Parker, Ph.D.
*Director, Clinical Data Programs*
*Healthcare Quality and Analysis Division*

## Study Consultants

**University of California, San Francisco**

R. Adams Dudley, M.D., M.B.A.
*Principal Investigator*

Michael Kuzniewicz, M.D., M.P.H.
*Co-investigator*

Mitzi Dean, M.S., M.H.A.
*Project Director*

Rondall Lane, M.D., M.P.H.
*Co-investigator*

Deborah Rennie
*Programmer/Analyst*

Peter Bacchetti, Ph.D.
*Statistician*

Ted Clay, M.S.
*Programmer/Analyst*

Stephen Crane, M.B.A.
*Research Associate*

Hal Luft, Ph.D.
*Co-investigator*

## California Health Policy and Data Advisory Commission


William Weil, M.D., Chair
*Representing Group Prepayment Health Plans*

Corinne Sanchez, Esq.
*General Member*

M. Bishop Bastien
*Representing Health Insurance Companies*

William Brien, M.D.
*Representing Physicians and Surgeons*

Paula Hertel, M.S.W.
*Representing Long-Term Care Facilities*

Janet Greenfield, R.N.
*Freestanding Ambulatory Surgery Centers*

Howard L. Harris, Ph.D.
*General Member*

Marjorie B. Fine, M.D., FACS
*General Member*

Sol Lizerbram, D.O.
*General Member*

Vacant
*Representing Disproportionate Share Hospitals*

Hugo Morris
*Representing Labor Health Coalitions*

Jerry Royer, M.D., M.B.A.
*Representing Hospitals*

Kenneth M. Tiratira, M.P.A.
*Business Health Coalitions*

Jacquelyn Paige
*Executive Director*


## California Health Policy and Data Advisory Commission Technical Advisory Committee


Jerry Royer, M.D., M.B.A., Chair
*Representing Hospitals*

Douglas Bagley
*Representing Hospitals*

Marilyn Chow, R.N., D.N.Sc.
*Representing California Nurses Association*

Mark Hlatky, M.D.
*Representing Research Community*

Diana Petitti, M.D.
*Representing Research Community*

Laurie Sobel, J.D.
*Representing Consumers Union*

Jeffrey Rideout, M.D.
*Representing California Medical Association*

Maida Reavis Herbst, R.H.I.A.
*Representing Medical Records Practitioner*

Laura B. Gardner, M.D., M.P.H., Ph.D.
*Representing Research Community*

David E. Hayes-Bautista, Ph.D.
*Representing Research Community*

Robert H. Brook, M.D., Sc.D.
*Representing Research Community*

Nancy Donaldson, R.N., D.N.Sc.
*Representing California Nurses Association*

William S. Weil, M.D.
*Representing California Medical Association*

## Index


| | *Page(s)* |
|---|---|
| 1. Executive Summary…… | 1-1 to 1-3 |
| 2. Literature Summary…… | 2-1 to 2-9 |
| 3. Sampling and Sample Size…… | 3-1 to 3-3 |
| 4. Patient Inclusion and Exclusion Criteria …… | 4-1 to 4-3 |
| 5. Hospital Selection and Patient Population…… | 5-1 to 5-4 |
| 6. Data Quality and Data Collection Process…… | 6-1 to 6-4 |
| 7. Model Descriptions and Development…… | 7-1 to 7-10 |
| 8. Customization of the Models…… | 8-1 to 8-3 |
| 9. Results I: Performance of the Original Models……..... | 9-1 to 9-4 |
| 10. Results II: Performance of the Customized Models…… | 10-1 to 10-8 |
| 11. Results III: Comparative Performance of the Hospitals…….. | 11-1 to 11-3 |
| 12. CALICO 2004…… | 12-1 to 12-3 |
| Appendix…… | A-1 to A-7 |

## Tables and Figures


| | *Page(s)* |
|---|---|
| Table 2.1 Performance of the models in the original datasets…….. | 2-4 |
| Table 2.2 Variables collected by each model……..…….. | 2-5 |
| Table 2.3 Summary of studies assessing performance of the models…… | 2-6 |
| Table 3.1 Width of confidence intervals on observed/expected mortality ratios.. | 3-2 |
| Table 5.1 Characteristics of hospitals included in the study……. | 5-2 |
| Table 5.2 Patients excluded from the CALICO database……..…… | 5-3 |
| Table 5.3 Patients excluded from the comparative analysis of the models……... | 5-3 |
| Table 5.4 Summary of patient population characteristics…….. | 5-4 |
| Table 6.1 Agreement statistics - training audit…….. | 6-2 |
| Table 6.2 Listing of additional variables in the CALICO data collection……. | 6-4 |
| Table 7.1 SAPS II variables and point assignments…….. | 7-3 |
| Table 7.2 Non-significant variables excluded from the SAPS II model…… | 7-3 |
| Table 7.3 Variables in the $MPM_0$II model and their estimated coefficients……. | 7-4 |
| Table 7.4 $MPM_0$ II variables and definitions…….…… | 7-5 |
| Table 7.5 The APACHE II severity of disease classification system…… | 7-7 |
| Table 7.6 APACHE II definitions of chronic severe organ insufficiency……. | 7-7 |
| Table 7.7 APACHE II weights for age and chronic health…… | 7-7 |
| Table 7.8 APACHE II diagnoses and coefficients……. | 7-8 |
| Table 8.1 Re-estimated models…….. | 8-2 |
| Table 9.1 Summary of discrimination and calibration – Original models……. | 9-1 |
| Figure 9.1 Calibration curves - Original models……...……... | 9-3 to 9-5 |
| Table 10.1 Summary of discrimination and calibration - Re-estimated models… (split #1)……...…. | 10-1 |
| Figure 10.1 Calibration curves - Re-estimated models……...……..... | 10-2 to 10-4 |

## Tables and Figures


| | *Page(s)* |
|---|---|
| Table 10.2 $MPM_0II$ re-estimated coefficients and odds ratios……... | 10-5 |
| Table 10.3 SAPS II re-estimated coefficients and odds ratios……... | 10-5 |
| Table 10.4 APACHE II re-estimated coefficients and odds ratios…… | 10-6 |
| Table 10.5 APACHE III re-estimated coefficients and odds ratios……... | 10.7 |
| Table 11.1 Statistically significant high and low hospital outliers…… | 11-1 |
| Table 11.2 Rank order of hospitals' SMRs……. | 11-2 |
| Table 11.3 SMR and 95% confidence interval of the hospitals……... | 11-3 |
| Table A.1 Hosmer-Lemeshow Goodness-of-Fit (C test) – original models……. | A-1 to A-2 |
| Table A.2 Hosmer-Lemeshow Goodness-of-Fit (H test) – original models……. | A-3 to A-4 |
| Table A.3 $MPM_0II$ variables and re-estimated coefficients (all 3 splits)……. | A-5 |
| Table A.4 SAPS II variables and re-estimated coefficients (all 3 splits)……. | A-5 |
| Table A.5 APACHE II variables and re-estimated coefficients (all 3 splits)….. | A-6 |
| Table A.6 APACHE III variables and re-estimated coefficients (all 3 splits)…. | A-7 |

## Executive Summary

***Background***

The California Hospital Outcomes Project (CHOP) is an initiative mandated by the State of California and conducted by the Office of Statewide Health Planning and Development (OSHPD) to produce public reports comparing hospital outcomes for patients treated in hospitals throughout California for selected conditions, procedures, and units. Intensive care unit (ICU) mortality was selected as a potential topic for outcome reporting by the OSHPD staff and the California Health Policy and Data Advisory Commission Technical Advisory Committee because of the high mortality rate in ICUs, evidence that intervention could effect beneficial change in mortality, and the high cost of care in these specialized units.

There are four ICU risk-adjustment models that are widely used, the Mortality Probability Model II at admission or "zero hours" ($MPM_0II$), the Simplified Acute Physiology Score II (SAPS II), and the Acute Physiology and Chronic Health Evaluation, versions II and III (APACHE II and APACHE III). The utility of these four models in assessing the performance of modern ICUs is unknown, as they were developed from the mid-1980s to the early 1990s. Furthermore, no attempt has been made to compare these models, all of which use data obtained from chart abstractions, to models using only data already available in discharge abstracts. The existing mortality models vary significantly in the data burden they would place on hospitals.

***Objectives***

The goal of the California Intensive Care Outcomes (CALICO) project is to assess the feasibility of, potential benefits from, and most efficient approach to ICU performance reporting in California. One objective of CALICO is to evaluate the performance of $MPM_0II$, SAPS II, APACHE II, and APACHE III by applying them to a modern database of California ICU patients, including customizing the models to the California dataset to improve their goodness-of-fit. A second objective is to develop an ICU mortality risk-adjustment model using OSHPD's Patient Discharge Database (PDD) plus a small number of clinical data elements obtained by chart abstraction. To determine whether there is variability in performance, and hence potential for improvement, another objective is to make preliminary estimates of the range of risk-adjusted ICU mortality performance using these various models. The final objective is to compare the available models in terms of their predictive performance versus the burden of data collection—considering both the number of variables used and the sources from which those data are likely to be obtained—to identify the most efficient model or combination of models to report ICU performance.

***Methods***

To achieve the project goals, demographic, clinical, and limited therapeutic data were collected on ICU patients from 23 California hospitals that volunteered to join CALICO. Each hospital was instructed to collect data on consecutive, eligible patients starting with discharges on March 1, 2002 and continuing until their target sample size was reached. Eligible patients were adults (18 or older) who were admitted for at least 4 hours into an adult ICU and who were not burn, trauma, or coronary bypass patients. Patients admitted to rule out myocardial infarction who were not found to have a critical illness were excluded. Data was subsequently received on

6,108 patients, 5,473 of whom met the inclusion criteria for the study. For comparison across the models, only patients for whom a risk estimation could be calculated for all four of the extant ICU mortality models were used. This resulted in a database of 4,630 patients to use for the comparative analysis of the models. Mortality predictions were calculated for each patient using the four extant models with the coefficients as published by their developers and after re-estimating the models (using the same variables but recalculating the coefficients) on a 60% development sub-sample of the CALICO data.

***Results: Model Performance - Discrimination and Calibration***

$MPM_0$ II had a significantly lower discrimination (area under the receiver operating characteristic curve or AUC) than the other models using both the original and re-estimated coefficients. The AUC for $MPM_0$ II was 0.80 using the original coefficients and 0.81 using the re-estimated coefficients. In comparison, the respective AUCs for SAPS II were 0.87 and 0.88, for APACHE II 0.85 and 0.86, and for APACHE III 0.88 and 0.88.

All the models in their original forms were poorly calibrated to 2002 California data. Their Hosmer-Lemeshow C and H statistics were all significant with P values $< 0.05$. However, re-estimated versions were all reasonably calibrated with non-significant Hosmer-Lemeshow statistics when applied to a validation sub sample.

***Results: Hospital Comparisons***

To further evaluate the feasibility of hospital performance reporting, we calculated relative hospital performance using the various models. Depending on the model used, 29-76% of the hospitals were identified as low mortality outliers using the original coefficients. Once the models were re-estimated, fewer hospitals were identified as outliers. The APACHE III model was the most conservative in predicting outliers (two outliers, one low and one high) compared to APACHE II (five outliers, two low and three high), $MPM_0$ II (six outliers, four low and two high), and SAPS II (seven outliers, three low and four high). One hospital was identified, as a low mortality outlier irrespective of which model was used and one hospital was a high outlier in all models.

In general, regardless of the model used, the same hospitals were consistently ranked in the same quartile. When hospitals changed quartiles across models, they generally moved only a single quartile, although two hospitals crossed two quartiles. No hospitals moved from the top quartile to the bottom quartile or from the bottom to the top. The four models' point estimates of standardized mortality ratios (SMR) for each hospital were not statistically different from each other for 15 of 17 hospitals that reported on at least 100 patients. However, there were significant variations in the SMRs (irrespective of the model used) among hospitals after risk adjustment.

***Conclusions***

The four risk-adjustment models with their original coefficients should not be used for California performance reporting as their calibration is poor, leading to significant over-estimation of predicted mortality and the inappropriate classification of a large number of hospitals as low

mortality outliers. However, all four risk-adjustment models, once they have been re-estimated on California data, may be practical for ICU mortality reporting in the state. The models differ in data burden and performance characteristics. $MPM_0$ II has statistically significantly lower discrimination than the other three models, but the others, especially APACHE III, require more data than $MPM_0$ II. Regardless of the model used, some hospitals seem to perform in ways that are demonstrably different from expected (with hospitals at both the high and low ends of the mortality performance spectrum), further illustrating the need for hospital performance reporting to identify these hospitals. The choice of model should be based on the preferred balance between data burden and model performance. Areas that require additional examination are risk models developed using administrative data from the PDD, further adaptations of the current models, and changes in model performance when using a larger database.

## Literature Summary

### Importance of ICU

Intensive care units became common in hospitals in the United States during the polio epidemics of the 1950's. With the development of mechanical ventilators to assist breathing, a separate, specialized unit within the hospital was needed. Further developments in medical technology necessitated the expansion of intensive care to facilitate the handling of other types of severe organ dysfunction.

The modern intensive care unit (ICU) is the highest mortality unit in any hospital. There are approximately 4 million ICU admissions per year in the United States with average mortality rates reported ranging from 8-19%, or about 500,000 deaths annually.[1,2,3,4] This mortality rate is higher than for any specific condition or procedure, with the exception of myocardial infarction, for which California hospital performance reports have been developed or are being considered. Studies have shown 2- to 3-fold variations in mortality rates in ICUs, after adjusting for patient risk factors.[5,6,7] Since the patient population is severely ill and undergoes multiple complex interventions at the same time, these patients are extremely vulnerable to experiencing adverse outcomes. The ICU is also one of the sites in which medical errors are most likely to occur because of the complexity of care.[8,9,10,11] The ICU patient population is more likely to experience poor outcomes secondary to errors because these critically ill patients have little "reserve." [12,13,14] In addition to its impact on mortality, the ICU is an expensive component of the national healthcare budget, accounting for approximately 10% of the total hospital budget.[15]

Based on the clinical significance of the ICU, the Joint Commission on Accreditation of Hospital Organizations (JCAHO), and the Leapfrog Group have decided that ICU care is a priority. The National Quality Forum and the Leapfrog Group are considering requiring or recommending ICU performance reporting. JCAHO has already announced its intention to make ICU risk-adjusted mortality a core measure of performance. Given the levels of mortality and apparent variations in performance, there may be much to be gained from public reporting of ICU outcomes. Hospitals and clinicians would have benchmarks to use in setting quality improvement goals. Consumers could incorporate the information into decisions about their choice of hospital and perhaps also health plan or primary physician. However, these potential benefits can only be realized if ICU performance can be assessed accurately.

### Extant ICU Mortality Models

Clinicians and researchers have long recognized how important ICU performance is to overall hospital mortality and a significant amount of work has already been done to develop tools to assess ICU performance. This work has focused primarily on the development of general predictive models to compare observed versus expected mortality rates across a wide range of patients. Disease-specific models and organ dysfunction/failure models have also been developed. The disease-specific models have the advantage of identifying specific variables for a particular disease that will affect outcome, theoretically improving the predictive power of a model compared to models developed for all ICU patients. However, studies have shown mixed results, in terms of improving discrimination, for disease-specific models versus general models.[16] Since disease-specific models only predict outcomes for a specific disorder, they

cannot be used for the majority of ICU patients. As a consequence, unlike the general models, they cannot be used to examine mortality of an entire ICU population or compare overall performance among ICUs.

Organ dysfunction/failure models were developed for the quantification of multiple organ dysfunction syndrome and have shown a good correlation between the presence and duration of organ failure and outcome.[17] These models are often relatively simple and are used to describe individual organ dysfunction/failure in a continuous form, from mild dysfunction to severe failure. They can be used over a period of time to monitor the progression of individual (or aggregated) organ dysfunction.[16] These models are intended to be used to describe morbidity and have not been shown to be accurate for mortality prediction. Consequently, general ICU mortality prediction models are currently the most effective in evaluating the performance of ICUs.

Four major general ICU mortality risk-adjustment models are widely used; the Acute Physiology and Chronic Health Evaluation, versions II and III (APACHE II and III), the Simplified Acute Physiology Score II (SAPS II) and the Mortality Probability Model II at admission or "zero hours" ($MPM_0$ II).[4,18,19,20] The models have been used in more than 2,000 publications in the medical literature. These second and third generation revised models represent an effort to improve the performance of the original models, which are no longer in wide use. The performance of these four models is summarized in Table 2.1. Chapter 7 details the specific steps in model generation, the variables selected, risk equations, and coefficients of variables.

These models were developed in the mid-1980's to early 1990's. The models were developed using multiple logistic regression equations. While the models do have some similar variables collected, they differ greatly in the number of variables collected, the type of variables, and specification of the variables. Table 2.2 lists the variables collected in each model. The difference in the number of variables collected and the difficulty in collecting the variables are significant because it has a major impact on the time required to abstract the variables and ultimately the cost of generating mortality predictions. The APACHE II model includes 12 physiologic variables, 6 chronic health variables, age, source of admission, and 50 reasons for admission categories. The revised APACHE III model contains 20 physiologic variables, 7 chronic health variables, 7 categories for age, source of admission, and 78 reasons for admission categories.[4] The $MPM_0$ II model and SAPS II models were developed by researchers in North America and Europe in an attempt to simplify ICU mortality prediction. SAPS II, widely used in France, uses 12 physiologic variables, age, a chronic health variable, and type of admission (surgical vs. medical).[19] $MPM_0$ II uses 10 physiologic variables, 3 chronic health variables, age and type of admission.[20] Neither the $MPM_0$ II model nor the SAPS II requires the data collector to determine a reason for admission to the ICU as the APACHE models do.
Since the development of the models, their developers have revised and expanded their work, although, with one exception, the risk adjustment models are still calculated in the same manner. The APACHE III model has added a coefficient for length of stay in the hospital after ICU discharge. Data collection to revise the SAPS model is currently ongoing in Europe (see www.saps3.org); the developers are assessing the possibility of adding a single reason for admission as well as the presence of infection at admission. However, most of their revisions are

aimed at developing a multi-dimensional model of ICU operations and investigating other outcomes and cost-effectiveness.

## Use of ICU Mortality Models for Performance Evaluation

Among the general ICU models, no model is obviously superior to the others for the purpose of ICU performance evaluation. The APACHE models are the most widely used, in part because their complexity gives them greater clinical plausibility. To compare the models, one must assess the calibration and discrimination of each. Discrimination is assessed with areas under the receiver operating characteristic curves (AUCs). The calibration is evaluated by Hosmer-Lemeshow statistics. Hosmer-Lemeshow statistics test whether one can reject the null hypothesis that the model fits well across deciles of risk, so that on average, people with high predicted values have comparable mortality rates, and contrariwise for those with low predicted values. A high p-value indicates a good fit. The relative importance of model calibration versus model discrimination depends on the intended use of the model. Both assessments are needed to identify a well-fit model. Hosmer and Lemeshow have argued that if a model does not calibrate well, it is meaningless to examine discrimination.[21] Calibration is the key measure if the model is anticipated to be used to compare predicted and actual death rates, and thus compare performance across hospitals, especially if the risk profile varies among hospitals.

All the general models seemed to perform well on the populations on which they were developed, although we were unable to find reports of the calibration of the APACHE models on the development data sets. For all models, AUCs were 0.82 or better on the developmental and validation samples. The $MPM_0$ II and SAPS II studies, using Hosmer-Lemeshow statistics, demonstrated good calibration ($MPM_0$ II model P= 0.623 and P= 0.327 in the developmental and validation samples respectively, SAPS II, P= 0.883 and P=0.104). However, when applied to populations other than the ones on which they were developed and validated, all four models discriminate adequately but calibrate poorly.[22,23,24,25,26,27,28,29,30,31,32,33,34] When comparing APACHE III to other models, especially APACHE II, in general the additional variables added to APACHE III that distinguish it from APACHE II do lead to increased discrimination. However, prior investigators have found that this discriminatory power comes at a cost of poorer calibration in most cases. Table 2.3 is a summary of studies assessing and comparing the calibration and discrimination of the models.

A number of hypotheses have been put forward to account for these findings, including differences in the definition and collection of the data, real differences in the patient populations (case-mix), lead time bias, lack of important predictive variables or interactions between the variables in the models, pre-ICU or post-ICU management, or a lack of validity of the dependent variable.[16,35,36] Some studies have shown that the models do not calibrate well because they underestimate the mortality of low-risk patients and overestimate the mortality of high-risk patients.[35] Other researchers have shown the problematic nature of utilizing the worst physiologic variables over the first 24 hours of an ICU admission, knowing that physiologic data can be strongly influenced by medical and nursing intervention.[37] As a result, patients treated inappropriately in the first 24 hours after admission may receive higher mortality scores, even when their risk of mortality at admission was lower. Lead time bias is an issue, as a large proportion of the patients admitted to an ICU come from the emergency department.

Consequently, differences in the treatment of individuals in the emergency department will affect the degree of a patient's physiologic derangements at ICU admission and thus their mortality score.[38] Therefore, hospitals that do a superior job stabilizing their patients in the emergency department may appear to have a case-mix with lower predicted mortality. Finally, critical care practice, technology and knowledge have changed significantly since the development of the newest models more than 10 years ago and the coefficients used in the models need to be reassessed to reflect modern practices and outcomes.

Approaches to resolving problems through model innovation, such as restricting mortality assessment to patients above a specific risk threshold, eliminating transfer patients, or assessing clinically defined subgroups from the assessed population, have not, to our knowledge, been attempted for ICU mortality models. Although some of the original models have been shown to have reasonable discrimination at the patient level and adequate calibration among deciles of patients, there is legitimate concern about their usefulness when comparing ICUs to each other without explicitly considering ICU thresholds for admission. Some hospitals may have estimated risks above or below their true risks because they use different admission criteria for their ICUs patients.

Highly predictive risk models have already been developed for pneumonia and myocardial infarction in Californian populations, as well as congestive heart failure (CHF) and pneumonia in national populations.[39,40,41,42,43,44,45] In reviewing the literature on condition-specific mortality risk models for conditions that have a high prevalence in the ICU and/or that result in high mortality, several key risk factors were not included in the general ICU mortality models. A model that includes some of these factors may perform better than any of the existing models.

## Table 2.1
## Performance of models in the original datasets

| Model | $MPM_0II$ | SAPS II | APACHE II | APACHE III |
|---|---|---|---|---|
| Year of Publication | 1993 | 1993 | 1985 | 1992 |
| Hospital Mortality | 20.8% | Varied from 13.8% in Switzerland to 32.4% in the UK | Varied by diagnosis | Varied from 0.9% for drug overdose to 65.9% for cardiogenic shock |
| Discrimination (AUC)* | 0.837 for developmental<br>0.824 for validation | 0.88 for developmental<br>0.86 for validation | 0.863 | 0.90 |
| Calibration (H-L Statistic)† | P = 0.623 developmental<br>P = 0.327 validation | P = 0.883 developmental<br>P = 0.104 validation | Not given | Not given |
| Data Reliability (% of agreement) | 96-99% for dichotomous variables, 63% for chronic renal insufficiency, 80-85% for other variables | 81% for potassium; >87% for other variables | 96% on physiologic variables; less on preadmission data | 90% on APS‡ score; 85.7 - 99.5% for other variables |

*= Area under the receiver operating characteristic curve †= Hosmer-Lemeshow Statistic; C test for MPM, H test for SAPS ‡ = Acute Physiology Score

## Table 2.2
## Variables collected by each model

| VARIABLES | $MPM_0$ II | SAPS II | APACHE II | APACHE III |
|---|---|---|---|---|
| **Chronic Health Status** | | | | |
| AIDS | | ✓ | ✓ | ✓ |
| Cirrhosis | ✓ | | ✓ | ✓ |
| Lymphoma | | | ✓ | ✓ |
| Hematologic malignancy | | ✓ | ✓ | |
| Leukemia | | | ✓ | ✓ |
| Hepatic Failure | | | ✓ | ✓ |
| Metastatic Cancer | ✓ | ✓ | ✓ | ✓ |
| Immunosuppression | | | ✓ | ✓ |
| Chronic Renal Insufficiency | ✓ | | ✓ | ✓ |
| **Physiology** | | | | |
| Temperature | | ✓ | ✓ | ✓ |
| Heart Rate | ✓ | ✓ | ✓ | ✓ |
| Respiratory Rate | | | ✓ | ✓ |
| Blood Pressure | ✓ | ✓ | ✓ | ✓ |
| White Blood Cell Count | | ✓ | ✓ | ✓ |
| Albumin | | | | ✓ |
| Bilirubin | | ✓ | | ✓ |
| Electrolytes | | ✓ | ✓ | ✓ |
| Blood Urea Nitrogen | | ✓ | | ✓ |
| Creatinine | | | ✓ | ✓ |
| Urine Output | | ✓ | | ✓ |
| Blood Gas | | ✓ | ✓ | ✓ |
| Glasgow Coma Score | ✓ | ✓ | ✓ | ✓ |
| **Acute Diagnoses** | | | | |
| Acute Renal Failure | ✓ | | | |
| Arrhythmias | ✓ | | | |
| Cerebrovascular Accident | ✓ | | | |
| GI Bleeding | ✓ | | | |
| Leukemia | ✓ | | | |
| Infection | | | | |
| Intracranial Mass Effect | ✓ | | | |
| Select one from a list of … | | | 50 diagnoses | 78 diagnoses |
| **Other** | | | | |
| Age | ✓ | ✓ | ✓ | ✓ |
| Patient origin | | | ✓ | ✓ |
| CPR prior to ICU admission | ✓ | | | |
| Mechanical Ventilation | ✓ | | | |
| Vasoactive Drug Therapy | | | | |

## Table 2.3
## Summary of studies assessing performance of the models

| Study Authors | Pub. Year | Location | Hospital Mortality | # of Patients | Model | AUCs* | H-L Statistic† | Test | P-value |
|---|---|---|---|---|---|---|---|---|---|
| Livingston et. al. | 2000 | Scotland | 29.40% | 10,334 | SAPS II | 0.78 | 142.0 | C | < 0.05 |
| | | | | 10,393 | MPM $_0$ II | 0.74 | 451.8 | C | < 0.05 |
| | | | | 9,848 | APACHE II | 0.76 | 67.4 | C | < 0.05 |
| | | | | 9,848 | UK APACHE II ‡ | 0.76 | 236.8 | C | < 0.05 |
| | | | | 10,326 | APACHE III | 0.80 | 365.7 | C | < 0.05 |
| Zimmerman et. al. | 1998 | USA | 12.35% | 36, 668 | APACHE III | 0.89 | 35.8 | C | < 0.0001 |
| | | | | | | | 48.7 | H | < 0.0001 |
| Moreno et. al. | 1998 | Europe | 20.00% | 10,027 | SAPS II | 0.82 | 208.4 | C | <0.0001 |
| | | | | | | | 218.2 | H | <0.0001 |
| | | | | | MPM $_0$ II | 0.79 | 368.2 | C | <0.0001 |
| | | | | | | | 437.1 | H | <0.0001 |
| Pappachan et. al. | 1999 | England | 25.90% | 12,793 | APACHE III | 0.89 | 332.9 | C | <0.01 |
| | | | | | | | 312.5 | H | <0.01 |
| Beck et. al. | 1997 | UK | 26.11% | 1,144 | APACHE II | 0.80 | 98.6 | C | <0.05 |
| | | | | | APACHE III | 0.85 | 129.8 | C | <0.05 |
| Markgraf et. al. | 2000 | Germany | 18.50% | 2,661 | SAPS II | 0.85 | 20.5 | C | <0.01 |
| | | | | 2,795 | APACHE II | 0.83 | 11.8 | C | >0.1 |
| | | | | 2,661 | APACHE III | 0.85 | 48.1 | C | <0.001 |
| Rowan et. al. | 1993 | Britain & Ireland | 27.70% | 8,796 | APACHE II | 0.83 | 79.8 | H | <0.05 |
| Castella | 1995 | Europe & N. America | 21.80% | 4,099 | SAPS II | 0.85 | n/a | C | 0.0244 |
| | | | | | | | n/a | H | 0.1019 |
| | | | | | MPM $_0$II | 0.81 | n/a | C | 0.072 |
| | | | | | | | n/a | H | 0.0148 |
| | | | | | APACHE II | 0.86 | n/a | C | 0.0245 |
| | | | | | | | n/a | H | 0.0074 |
| | | | | | APACHE III | 0.86 | n/a | C | n/a |
| | | | | | | | n/a | H | n/a |
| Apolone et. al. | 1996 | Italy | 34% | 1,393 | SAPS II | 0.8 | 71.0 | H | <0.001 |
| Moreno et. al. | 1997 | Portugal | 32% | 982 | SAPS II | 0.82 | 28.3 | C | 0.002 |
| | | | | | | | 29.7 | H | 0.001 |
| | | | | | APACHE II | 0.79 | 49.7 | C | <0.0001 |
| | | | | | | | 32.7 | H | 0.0003 |
| Metnitz et. al. | 1999 | Austria | 19.50% | 1,733 | SAPS II | 0.81 | 91.8 | C | <0.0001 |
| | | | | | | | 89.1 | H | <0.0001 |
| Bastos et. al. | 1996 | Brazil | 34% | 1,734 | APACHE II | 0.82 | n/a | n/a | n/a |
| Rivera-Fernandez | 1998 | Spain | 21.10% | 10,929 | APACHE III ‡ | 0.83 | 12.27 | H | >0.5 |

* = Area under the receiver operating characteristic curve † = Hosmer-Lemeshow statistic ; C Test or H Test
‡ - these models were modified by having the coefficients for the variables re-estimated based on the study population

## References:


[1] Wu AW, Pronovost P, Morlock L. ICU incident reporting systems. *J Crit Care.* Jun 2002;17(2):86-94.

[2] Angus DC, Linde-Zwirble WT, Sirio CA, et al. The effect of managed care on ICU length of stay: implications for Medicare. *JAMA.* Oct 2 1996;276(13):1075-1082.

[3] Young MP, Birkmeyer JD. Potential reduction in mortality rates using an intensivist model to manage intensive care units. *Eff Clin Pract.* Nov-Dec 2000;3(6):284-289.

[4] Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest.* Dec.1991;100(6):1619-1636.

[5] Gunning K, Rowan K. ABC of intensive care: outcome data and scoring systems. *BMJ.* Jul 24 1999;319(7204):241-244.

[6] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. An evaluation of outcome from intensive care in major medical centers. *Ann Intern Med.* Mar 1986;104(3):410-418.

[7] Shortell SM, Zimmerman JE, Gillies RR, et al. Continuously improving patient care: practical lessons and an assessment tool from the National ICU Study. *QRB Qual Rev Bull.* May 1992;18(5):150-155.

[8] Brennan TA, Leape LL, Laird NM, et al. Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Engl J Med*. Feb 7 1991;324(6):370-376.

[9] Cullen DJ, Sweitzer BJ, Bates DW, Burdick E, Edmondson A, Leape LL. Preventable adverse drug events in hospitalized patients: a comparative study of intensive care and general care units. *Crit Care Med.* Aug 1997;25(8):1289-1297.

[10] Thomas EJ, Studdert DM, Burstin HR, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. *Med Care.* Mar 2000;38(3):261-271.

[11] Andrews LB, Stocking C, Krizek T, et al. An alternative strategy for studying adverse events in medical care. *Lancet.* Feb 1 1997;349(9048):309-313.

[12] Giraud T, Dhainaut JF, Vaxelaire JF, et al. Iatrogenic complications in adult intensive care units: a prospective two-center study. *Crit Care Med.* Jan 1993;21(1):40-51.

[13] Pronovost P, Wu AW, Dorman T, Morlock L. Building safety into ICU care. *J Crit Care.* Jun 2002;17(2):78-85.

[14] Joint Commission on Accreditation of Healthcare Organizations. What every hospital should know about sentinel events. Oakbrook Terrace, IL: *JCAHO*; 2000.

[15] Wagner DP, Wineland TD, Knaus WA. The hidden costs of treating severely ill patients: charges and resource consumption in an intensive care unit. *Health Care Financ Rev.* Fall 1983;5(1):81-86.

[16] Moreno R. The "new" scores: what problems have been fixed, and remain? *Current Opinion in Critical Care.* 2000;6:158-165.

[17] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. Prognosis in acute organ-system failure. *Ann Surg.* Dec 1985;202(6):685-693.

[18] Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med.* Oct 1985;13(10):818-829.

[19] Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA.* Dec 22-29 1993;270(24):2957-2963.

[20] Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA.* Nov 24 1993;270(20):2478-2486.

[21] Mourouga P GC, Rowan KM. Does it fit? Assessment of scoring systems. *Current Opinion in Critical Care.* 2000;6:176-180.

[22] Apolone G, Bertolini G, D'Amico R, et al. The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva. *Intensive Care Med.* Dec 1996;22(12):1368-1378.

[23] Moreno R, Morais P. Outcome prediction in intensive care: results of a prospective, multicentre, Portuguese study. *Intensive Care Med.* Feb 1997;23(2):177-186.

[24] Moreno R, Reis Miranda D, Fidler V, Van Schilfgaarde R. Evaluation of two outcome predictors on an independent database. *Crit Care Med.* 1998;26:50-61.

[25] Metnitz PG, Valentin A, Vesely H, et al. Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Simplified Acute Physiology Score. *Intensive Care Med.* Feb 1999;25(2):192-197.

[26] Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland--II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. *BMJ.* Oct 16 1993;307(6910):977-981.

[27] Bastos PG, Sun X, Wagner DP, Knaus WA, Zimmerman JE. Application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. *Intensive Care Med.* Jun 1996;22(6):564-570.

[28] Rivera-Fernandez R, Vazquez-Mata G, Bravo M, et al. The Apache III prognostic system: customized mortality predictions for Spanish ICU patients. *Intensive Care Med.* Jun 1998;24(6):574-581.

[29] Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA. Evaluation of acute physiology and chronic health evaluation III predictions of hospital mortality in an independent database. *Crit Care Med.* Aug 1998;26(8):1317-1326.

[30] Castella X, Artigas A, Bion J, Kari A. A comparison of severity of illness scoring systems for intensive care unit patients: results of a multicenter, multinational study. The European/North American Severity Study Group. *Crit Care Med.* Aug 1995;23(8):1327-1335.

[31] Livingston BM, MacKirdy FN, Howie JC, Jones R, Norrie JD. Assessment of the performance of five intensive care scoring models within a large Scottish database. *Crit Care Med.* Jun 2000;28(6):1820-1827.

[32] Markgraf R, Deutschinoff G, Pientka L, Scholten T. Comparison of acute physiology and chronic health evaluations II and III and simplified acute physiology score II: a prospective cohort study evaluating these methods to predict outcome in a German interdisciplinary intensive care unit. *Crit Care Med.* Jan 2000;28(1):26-33.

[33] Beck DH, Taylor BL, Millar B, Smith GB. Prediction of outcome from intensive care: a prospective cohort study comparing Acute Physiology and Chronic Health Evaluation II and III prognostic systems in a United Kingdom intensive care unit. *Crit Care Med.* Jan 1997;25(1):9-15.

[34] Pappachan JV, Millar B, Bennett ED, Smith GB. Comparison of outcome from intensive care admission after adjustment for case mix by the APACHE III prognostic system. *Chest.* Mar 1999;115(3):802-810.

[35] Beck DH, Smith GB, Taylor BL. The impact of low-risk intensive care unit admissions on mortality probabilities by SAPS II, APACHE II and APACHE III. *Anaesthesia.* Jan 2002;57(1):21-26.

[36] Glance LG, Osler TM, Papadakos P. Effect of mortality rate on the performance of the Acute Physiology and Chronic Health Evaluation II: a simulation study. *Crit Care Med.* Oct 2000;28(10):3424-3428.

[37] Boyd O, Grounds RM. Physiological scoring systems and audit. *Lancet.* Jun 19 1993;341(8860):1573-1574.

[38] Tunnel RD, Millar BW, Smith GB. The effect of lead time bias on severity of illness scoring, mortality prediction and standardized mortality ration in intensive care – a pilot study. *Anaesthesia.* 1998;53:1045-1053.

[39] Haas J, Romano P, Dean ML, Hung Y, Bacchetti P. Report for the California Hospital Outcomes Project, Community-Acquired Pneumonia, 1996. Sacramento, California: Office of Statewide Planning and Development, November 2000.

[40] Haas J, Dean ML, Hung Y, Rennie DJ. Differences in mortality among patients with community-acquired pneumonia in California by ethnicity and hospital characteristics. *Am J Med.* Jun 1 2003;114(8):660-664.

[41] Luft HS, Romano PS. Report on Heart Attack, 1991-1993, Volume 3: Detailed Statistical Results. Sacramento, CA. California Office of Statewide Health Planning and Development, December 1997.

[42] Rosenthal GE, Harper DL, Shah A, Covinsky KE. A regional evaluation of variation in low-severity hospital admissions. *J Gen Intern Med.* Jul 1997;12(7):416-422.

[43] Alexander M, Grumbach K, Selby J, Brown AF, Washington E. Hospitalization for congestive heart failure. Explaining racial differences. *JAMA.* Oct 4 1995;274(13):1037-1042.

[44] Fine MJ, Orloff JJ, Arisumi D, et al. Prognosis of patients hospitalized with community-acquired pneumonia. *Am J Med.* May 1990;88(5N):1N-8N.

[45] LaCroix AZ, Lipson S, Miles TP, White L. Prospective study of pneumonia hospitalizations and mortality of U.S. older people: the role of chronic conditions, health behaviors, and nutritional status. *Public Health Rep.* Jul-Aug 1989;104(4):350-360.

# Sampling and Sample Size

## Hospital and Patient Sampling

The hospital sample is a voluntary sample, stratified by size. Although the sample is voluntary, an extraordinary effort was made to recruit a cross section of hospitals to include teaching and non-teaching, urban and rural, with differing governance.

Hospitals collected data on consecutive, eligible patients who were discharged from the hospital after having an eligible stay in the ICU. Hospitals with up to 1,200 ICU admissions per year were asked to provide data on 200 patients, hospitals with between 1,200 and 2,400 admissions per year provided data on 400 patients and hospitals with more than 2,400 ICU admissions per year provided data on 600 patients. Patient exclusions are described in detail in the next chapter.

Twenty-four hospitals participated in the 2002-2003 data collection. They were stratified by size as follows: 3 hospitals with >2,400 ICU admissions a year, 6 hospitals with 1,200-2400 admissions, and 15 with 1,200 or fewer admissions. Six hospitals provided less than 100 patient records that could be used across all models and were not included in hospital comparative analyses. One >2,400 patient hospital reported cases too late for these analyses. The breakdown for hospitals used in analyses for this report by number of cases submitted and hospital size is provided in Table 3.1. The total sample for all project years will include hospitals that collected data in 2002-2003, and a second group of hospitals that includes many of the first year hospitals as well as new hospitals that are collecting data from August 2003 through July 2004.

## Sample Size

Statistical concerns led us to use a minimum targeted sample size of at least 200 patients per hospital. Below this level, confidence intervals on the observed over expected mortality rate at the hospital level can get wide. If all the hospitals contributed only 200 patients, we would need more than 80 hospitals to achieve a sample size similar to that used by APACHE. Furthermore, larger hospitals are likely to have more heterogeneity in patient populations and may be more likely to have concentrations of unusual patients (e.g., organ transplant patients, oncology patients). We needed to analyze different concentrations of patient populations and perhaps adjust for those differences to evaluate their effect on calibration and to be sure they are not biasing estimates of a hospital's performance. Addressing these concerns required more than 200 patients from larger hospitals. We adopted a graded sample size requirement in order to explore the effect of severity and case-mix differences at larger hospitals and to avoid unduly burdening smaller hospitals. To replicate the methodology used by APACHE and the other models, data from patients consecutively admitted to the ICU was collected.

The assessment of the width of the confidence intervals for the 2002-03 data collection stratified by size is reported in Table 3.1. Hospitals that were asked to report results on 600 patients had confidence interval widths in the .24-.29 range (this varies as a function of the observed over expected mortality ratio, or O/E). Hospitals that were to collect data on 400 patients had a confidence interval width of .29-.41 range, and those collecting data for 200 patients had confidence intervals between .40-.59. While smaller differences in O/E would be clinically

meaningful, these confidence interval widths are within the range of confidence intervals for other public reporting projects.[1]

## Table 3.1
## Width of confidence intervals on observed/expected mortality ratios
**(O/Es using APACHE II, developer coefficients)**

### Hospitals requested to collect data on 600 patients

| Hospital | N* | Observed Deaths | Expected Deaths | O/E | 95% CI | Width of CI |
|---|---|---|---|---|---|---|
| Hospital P | 411 | 87 | 125.4 | 0.69 | 0.571-0.81 | 0.24 |
| Hospital E | 311 | 44 | 90.6 | 0.49 | 0.34-0.63 | 0.29 |

### Hospitals requested to collect data on 400 patients

| Hospital | N* | Observed Deaths | Expected Deaths | O/E | 95% CI | Width of CI |
|---|---|---|---|---|---|---|
| Hospital J | 320 | 54 | 63.8 | 0.85 | 0.65-1.04 | 0.39 |
| Hospital K | 369 | 61 | 84.8 | 0.72 | 0.56-0.88 | 0.32 |
| Hospital D | 326 | 52 | 88.5 | 0.59 | 0.44-0.73 | 0.29 |
| Hospital F | 328 | 37 | 57.4 | 0.65 | 0.43-0.73 | 0.30 |
| Hospital C | 355 | 37 | 70.8 | 0.52 | 0.34-0.71 | 0.37 |
| Hospital L | 268 | 32 | 55.1 | 0.58 | 0.38-0.78 | 0.41 |

### Hospitals requested to collect data on 200 patients

| Hospital | N* | Observed Deaths | Expected Deaths | O/E | 95% CI | Width of CI |
|---|---|---|---|---|---|---|
| Hospital I | 213 | 38 | 49.9 | 0.76 | 0.55-0.98 | 0.43 |
| Hospital H | 188 | 22 | 38.3 | 0.58 | 0.33-0.82 | 0.49 |
| Hospital N | 177 | 36 | 42.6 | 0.85 | 0.62-1.07 | 0.45 |
| Hospital M | 167 | 25 | 35.6 | 0.70 | 0.45-0.96 | 0.51 |
| Hospital O | 196 | 39 | 46.8 | 0.83 | 0.62-1.05 | 0.43 |
| Hospital G | 170 | 22 | 34.5 | 0.64 | 0.38-0.89 | 0.51 |
| Hospital B | 176 | 16 | 49.7 | 0.32 | 0.12-0.52 | 0.40 |
| Hospital A | 120 | 13 | 26.1 | 0.50 | 0.21-0.79 | 0.59 |
| Hospital Q | 148 | 49 | 33.8 | 1.45 | 1.22-1.68 | 0.45 |

** This is the number of cases that could be used across all models per hospital. The actual number of cases submitted per hospital is higher.*

## Power Calculations and Identifying Outliers

From the data in Table 3.1, hospitals that reported at least 100 patients and are targeting a total of 200 patients have a mean O/E mortality rate, using the APACHE II model, of 0.74, and the mean over all hospitals is 0.70. The largest standard error of O/E in Table 3.1 is 0.15 for Hospital A; this would be expected to decrease to 0.12 with 200 patients, and would decrease further if O/E were higher than average (because more observed deaths imply better precision). Prior literature shows mortality rates varying as much as two- and three-fold, after risk adjustment. Therefore, we calculated the power we would need to detect an outlier hospital whose true O/E mortality rate was twice the current mean of 0.70, or 1.40. We determined if the hospital collected data on 200 patients, we would have more than 99% power to detect such a hospital as being an outlier. In addition, if we looked only for hospitals that were 50% above mean O/E, or 1.05, we would have power of at least 84% to identify them with 200 cases eligible for analysis at the hospital. Therefore, we believe we will have reasonable power to identify outliers.

## References:

[1] Fink AS, Campbell DA Jr, Mentzer RM Jr, et al. The National Surgical Quality Improvement Program in non-Veterans Administration hospitals: initial demonstration of feasibility. *Ann Surg.* 2002; 236(3): 344-353.

## Patient Inclusion and Exclusion Criteria

Each hospital collected information on consecutive patients admitted to their intensive care unit. The intent of the study was to evaluate the performance of existing ICU risk-adjusted mortality prediction models. As a result, the inclusion and exclusion criteria reflect the parameters already established by these prior models. Patients were excluded from the analysis of a model if they did not have all the required data elements to calculate a mortality score for that model. Analysis was carried out both by comparing the models using an identical case-mix (patients who met inclusion criteria for all four models) and evaluating the models separately using all patients that met inclusion criteria for each particular model, regardless of their inclusion/exclusion status in the other models.

**Inclusion criteria for all models:**

1. **Age 18 or older**
   The study included adults only. The APACHE models were developed on a patient population ≥16 years-old.[1,2] SAPS II[3] and MPM II[4] were developed on a population ≥ 18 years old. The clinical spectrum of diseases for children is significantly different and would require recalibration of the existing models.

2. **1:1 or 2:1 patient: RN staffing**
   ICU care requires 1:1 or 2:1, patient: RN staffing. Secondary to bed availability issues, patients sometimes "board" in the ICU, where they are physically in the ICU but do not require ICU care. For these patients, the nursing to patient ratio is typically greater than 2:1. The study's purpose is to evaluate how these models predict ICU performance; so only patients requiring ICU care were included. Consequently, we excluded patients admitted to an ICU room with a ratio of more than two patients per RN.

3. **Admitted to an adult ICU**
   Patients admitted to pediatric ICUs may have significantly different risk of mortality for a given condition compared to adult ICUs due to the vastly different spectrum of disease and clinical expertise.

4. **Stay in the ICU for at least four hours**
   The outcomes of individuals admitted to the ICU for less then four hours often reflect the care prior to the ICU admission. Such short stays are usually ended either by death (often reflecting irresolvable problems prior to admission) or transfer, to another unit (often reflecting a change in the patient's clinical status). As a result, these individuals should not be used to evaluate ICU performance amongst hospitals.

**Exclusion criteria for all models:**

1. **Burn patients**
   Burn patients were excluded from the SAPS II, MPM II, and APACHE model development populations. Physiologic and clinical variables to predict mortality in burn patients are considerably different than those used to predict mortality in a general ICU population. Often these patients are treated in separate, specialized units. Furthermore, specific prognostic systems have been previously developed for this subset of patients.

2. **Trauma patients**
   Trauma patients were excluded from the CALICO data collection - even though they were included in the original SAPS II, MPM II, and APACHE model development populations. Currently, in most parts of the United States, trauma patients who are critically ill go to designated regional trauma centers. Thus, those centers would have trauma patients but no other hospitals in the region would. Since the goal of public performance reports is to allow consumers and others to compare hospitals on their treatment of similar groups of patients, it seemed inappropriate to include trauma patients. Furthermore, specific prognostic systems have been previously developed for trauma patients and would be more useful for assessing the performance of regional trauma centers (if this was desired) than general ICU models.

3. **Coronary artery bypass grafting surgery (CABG)**
   CABG patients represent a specialized group whose physiologic derangements do not predict the same risk of mortality as other patients in the ICU. California already has a public reporting system, and the CABG-specific risk adjustment model OSHPD uses in these reports is likely to have better predictive power than any general ICU model. Therefore, these patients were excluded from CALICO.

4. **Patients admitted to rule out myocardial infarction that are found within 24 hours of ICU admission to *not* have a myocardial infarction or another critical illness**
   Individuals who "rule out" for myocardial infarction (MI) essentially are admitted to the ICU for monitoring of chest pain or a similar symptom. When this symptom is not due to myocardial ischemia (or another accepted reason for ICU admission, such as rupture of a thoracic aortic aneurysm), their risk of death is close to zero. Thus, variation in hospital policies about what percentage of patients are admitted to rule out for MI could have a large influence on calculated performance (hospitals that admitted many such patients would have lower than predicted mortality). Since such policies are known to vary and could significantly affect performance, we excluded rule out MI patients from the model if they did not have an MI or other critical illness.

5. **Readmissions**
   Readmissions to the ICU during the same stay were collected in the 2002-2003 dataset but not used for modeling since interventions during the first admission may impact the patient's risk of mortality in the second admission. Patients who were readmitted to the ICU were not collected in the 2003-2004 dataset.

**MPM II and SAPS II Exclusion Criteria:**

1. **Cardiac Surgery**
   In both MPM and SAPS, the developmental and validation datasets excluded cardiac surgery patients. In APACHE III, cardiac surgery patients are included but have a separate prediction model that has not been released to the public.

## References:


[1] Knaus WA, Draper EA, Wagner DP, Zimmerman J. APACHE II: A severity of disease classification system. *Crit Care Med.* 1985; 13:818-829.

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

[3] Le Gall J, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA*. Dec 22-29 1993;270(24):2957-2963.

[4] Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA.* 1993; 270:2478-2486.

# Hospital Selection and Patient Population

## Hospital Selection

All hospitals with an eligible ICU in California were sent a recruitment packet including support letters from OSHPD, a letter from the project Principal Investigator, and materials describing the project for both the 2002-03 and 2003-04 data collection efforts. In addition, conference calls and presentations were made at the hospital level and at various ICU-related meetings. Follow-up materials further explaining the CALICO project were sent to hospitals that expressed interest after each hospital that received a mailing was contacted by telephone. The Project Director and a CALICO staff member who is a registered nurse called each potential ICU participant. If there was interest in the project, the Principal Investigator participated in a conference call with the appropriate hospital staff, including the ICU physician in charge, ICU nurse managers and quality improvement staff from the hospital. In addition, particular attention was paid to the recruitment of hospital systems, as decision-making is more complex at these institutions. Recruitment was done at the corporate level, and in the quality assurance and research departments of these institutions. A summary of the characteristics of the 23 hospitals with data included in the 2002-2003 analysis is found in Table 5.1.

## Patient population

The total sample comprised 6,108 patients. We excluded 25 patients for being under 18 or having a missing age; 194 patients because they were burn, trauma, or CABG patients; 200 patients because they were readmissions to the ICU; 51 patients due to missing data on status at hospital discharge; 95 patients who had duplicate records; and 70 patients whose ICU admission was less than 4 hours (summarized in Table 5.2) This left a database of 5,473 patients, from 23 different hospitals. To minimize data collection on patients ineligible for the study, additional checks have been built into the data collection software for 2003. In order to compare the performance of the models, patients were excluded if they were missing data required to calculate a mortality prediction for any of the models or if there was specific exclusion for one of the models. Consequently, an additional 843 patients were excluded (345 secondary to model specific exclusion criteria and 498 secondary to missing data), leaving 4,630 patients to use for comparison of the models. The additional exclusions are summarized in Table 5.3. The SAPS model requires a $FiO_2/PaO_2$ ratio for all ventilated patients; however, not all ventilated patients undergo a blood gas analysis in the first 24 hours of their ICU admission, either because they are rapidly extubated after ICU admission or have had a blood gas evaluation just prior to admission. As a result, the 137 patients excluded for missing this variable may not reflect entirely data that was not ascertained, but rather the fact that the blood gas evaluation was not clinically warranted in these circumstances.

The mean hospital mortality in the population was 15%. The mean age of patients was 62.2 years (standard deviation 17.35 years) with a median of 64 years and mode of 74 years. The youngest patient was 18 and the oldest 101. Approximately 53% of the patients were male. Seventy-eight percent of the population were admitted for a medical reason, while 18% were admitted for elective surgery and another 4% for emergency surgery. A greater percentage of the patients 65 years of age or older died. Similarly, a greater percentage of medical patients died. A

summary of the characteristics of the population by vital status at hospital discharge is presented in Table 5.4.

## Table 5.1
## Characteristic of hospitals included in the study

| Hospital Beds | ICU Beds | Type of ICU | Type of Hospital |
|---|---|---|---|
| 500-600 | 25-30 | Medical, Surgical & Coronary Care separate | urban, public, med school affiliation w/residency program |
| 200-300 | 20-25 | Medical, Surgical & Coronary Care separate | suburban, non-profit, community based, no med school affiliation |
| 500-600 | 30-35 | Medical/Surgical/Coronary combined | suburban, non-profit, med school affiliation w/residency program |
| 400-600 | 30-35 | Medical/Surgical/Coronary combined | suburban, non-profit, med school affiliation w/residency program |
| 100-200 | 15-20 | Medical/Surgical/Coronary combined | urban, non-profit, community based, no med school affiliation |
| 200-300 | 15-20 | Medical/Surgical/Coronary combined | urban, public, med school affiliation w/no residency program |
| 50-100 | 5-10 | Medical/Surgical/Coronary combined | urban, for profit, community based, no med school affiliation |
| 300-400 | 35-40 | Medical, Surgical & Coronary Care separate | urban, private, non-profit, med school affiliation w/no residency program |
| 300-400 | 20-25 | Medical/Surgical/Coronary Care/Neurological combined | suburban, non-profit, community based, no med school affiliation |
| 100-200 | 10-15 | Medical/Surgical/Coronary combined | suburban, non-profit, community based, no med school affiliation |
| 200-300 | 10-15 | Medical, Surgical & Coronary Care separate | suburban, non-profit, community based, no med school affiliation |
| 800-900 | 55-60 | Medical, Surgical, Coronary Care & Neurological separate | urban, public, med school affiliation w/residency program |
| 100-200 | 10-15 | Medical/Surgical/Coronary combined | rural, for profit, community based, no med school affiliation |
| 100-200 | 10-15 | Medical/Surgical/Coronary combined | suburban, non-profit, community based, no med school affiliation |
| 400-500 | 35-40 | Medical, Surgical & Coronary Care separate | urban, non-profit, med school affiliation w/residency program |
| 200-300 | 15-20 | Medical/Surgical/Coronary combined | suburban, public, no med school affiliation, residency program |
| 400-500 | 20-25 | Medical/Surgical/Coronary combined | rural, non-profit, community based, no med school affiliation |
| 200-300 | 5-10 | Medical/Surgical/Coronary combined | urban, for profit, community based, no med school affiliation |
| 100-200 | 5-10 | Medical/Surgical/Coronary combined | suburban, non-profit, community based, no med school affiliation |
| 100-200 | 5-10 | Medical/Surgical/Coronary combined | rural, non-profit, community based, no med school affiliation |
| 100-200 | 5-10 | Medical/Surgical/Coronary combined | rural, non-profit, community based, minor med school affiliation |
| 300-400 | 30-35 | Medical/Surgical/Coronary combined | suburban, non-profit, community based, no med school affiliation |
| 200-300 | 25-30 | Medical, Surgical & Coronary Care separate | suburban, non-profit, community based, no med school affiliation |

## Table 5.2
## Patients excluded from the CALICO database

| | |
|---|---|
| **Total patient charts abstracted** | **6,108** |
| **Reason for exclusion** | |
| Age less than 18, missing age | **25** |
| CABG, burn, and trauma patients | **194** |
| Duplicate Records | **95** |
| Readmissions to the ICU | **200** |
| ICU admission less than 4 hours | **70** |
| Missing outcome variable (alive or dead at hospital discharge) | **51** |
| **Total exclusions** | **635** |
| **Total patients in CALICO database** | **5,473** |

## Table 5.3
## Patients excluded from the comparative analysis of the models

| | |
|---|---|
| **Total patients in the CALICO database** | **5,473** |
| **Reason for exclusion*** | |
| ICU stay less than 8 hours (APACHE II) | **119** |
| Missing APACHE II diagnosis | **186** |
| Missing APACHE III diagnosis | **260** |
| Cardiac Surgery ($MPM_0$ II and SAPS) | **269** |
| Missing $FiO_2/PaO_2$ ratio (SAPS) | **137** |
| Missing pre-ICU length of stay (APACHE III) | **76** |
| **Total exclusions** | **1,041 (843†)** |
| **Total patients used for comparative analysis of the models** | **4,630** |

*** = Exclusion criteria listed are not mutually exclusive**
**† = 843 total patients were excluded from the final analysis, some patients had multiple exclusions**

## Table 5.4
## Summary of patient population characteristics

| | All Patients | | Deaths | | Survivors | |
|---|---|---|---|---|---|---|
| | N | (%) | N | (%) | N | (%) |
| **Age** | | | | | | |
| Age <65 | 2345 | (50.6) | 238 | (33.5) | 2107 | (53.8) |
| Age ≥ 65 | 2285 | (49.4) | 472 | (66.5) | 1813 | (46.2) |
| **Sex** | | | | | | |
| Female | 2179 | (47.1) | 347 | (48.9) | 1832 | (46.7) |
| Male | 2449 | (52.9) | 363 | (51.1) | 2086 | (53.2) |
| Missing | 2 | (<0.1)) | | | 2 | (0.1) |
| **Type of Patient** | | | | | | |
| Elective surgery * | 856 | (18.5) | 40 | (5.6) | 816 | (20.8) |
| Emergency surgery | 171 | (3.7) | 26 | (3.7) | 145 | (3.7) |
| Medical | 3603 | (77.8) | 644 | (90.7) | 2959 | (75.5) |
| **Location Prior to Admission** | | | | | | |
| Unknown | 53 | (1.1) | 8 | (1.1) | 45 | (1.1) |
| This hospital | 4237 | (91.5) | 647 | (91.1) | 3590 | (91.6) |
| Another hospital | 236 | (5.1) | 41 | (5.8) | 195 | (5.0) |
| Skilled nursing facility | 17 | (0.4) | 5 | (0.7) | 12 | (0.3) |
| Home | 13 | (0.3) | 0 | (0.0) | 13 | (0.3) |
| Ambulatory surgery center | 6 | (0.1) | 1 | (0.1) | 5 | (0.1) |
| Other | 68 | (154) | 8 | (1.1) | 60 | (1.5) |
| **If admitted from "This Hospital," location within the hospital admitted to ICU from** | | | | | | |
| Unknown | 18 | (0.4) | 1 | (0.2) | 17 | (0.5) |
| Emergency Department | 2269 | (53.6) | 348 | (53.8) | 1921 | (53.5) |
| Operating room/ recovery room | 1013 | (23.9) | 64 | (9.9) | 949 | (26.4) |
| Another ICU | 5 | (0.1) | 0 | (0.0) | 5 | (0.1) |
| Other Inpatient Care | 932 | (22.0) | 234 | (36.1) | 700 | (19.5) |

** elective surgery as defined by APACHE II*

# Data Quality and Data Collection Process

## Training and Data Quality

For each hospital involved in the CALICO project, at least one data collector (most hospitals sent more than one) was required to attend a day-long training session. The training session involved an overview of the project, an extensive didactic portion describing each of the data elements collected, and information intended to promote data quality. In addition, data collectors were trained on the use of the data entry software developed specifically for the CALICO project. Incorporated into the software are automated checks on the quality of data, including alerts for unexpected or impossible values for physiologic variables. Each data collector was required to perform data abstraction on sample medical records and submit their results to CALICO to demonstrate their proficiency with the data collection process before beginning actual chart abstraction.

After each hospital completed abstraction of their first 40 charts, an on-site data audit for training purposes was performed on 25% of their records. Ten charts at each hospital were randomly chosen, and external auditors hired by CALICO re-abstracted the variables for the CALICO project. After the results of this data collection were compared to the original data collection, any discrepancies between the two chart abstractions were discussed with the data collectors at the individual hospitals. In addition, the data collectors were re-educated on any areas of the data collection process in which they appeared to be having difficulty.

With the above procedures, we have achieved excellent data quality so far for most variables. All individual risk factors except reason for admission can be categorized into 2-9 categories, depending on how they are coded in the mortality model scores. For the physiological variables, kappas range from 0.79 to 1.0 for hospitals audited to this point. The median is 0.92, the 10th percentile 0.83 and the 90th percentile 0.98. Percent of exact agreement between two abstractors of the same chart was similarly good (see Table 6.1). Similarly good data quality was found for type of admission (medical vs. elective surgical vs. emergency surgical). While it is possible that data quality may decline as collectors get further from training, we attempt to maintain data quality through newsletters to data collectors describing difficult cases and how the investigators would score them, reporting patterns of disagreement, etc.

There are two problem areas for data quality. The first is the determination of which patients have chronic organ insufficiency. Although the percent agreement here was high (78%), the kappa was low (0.29). This was because most patients did not have any chronic organ insufficiency and the agreement between auditors about those few who might have had it was poor. The most problematic variable (not included in the kappas above because it cannot be categorized) is the reason for admission. The percent of exact agreement between auditors and hospital data collectors is only 64%. In most cases, the disagreements do not change predicted mortality probability because the conditions are grouped together in the APACHE reason for admission categories; however, in 27% of cases they change predicted mortality. These disagreements seem to reflect ambiguity in the clinical reality (e.g. Is the reason for admission

the malignant hypertension or the MI that has resulted?). We did allow data collectors to code a secondary reason for admission and will consider using this information in subsequent analyses.

The developers of the four models did not publish percent inter-rater agreement and kappas for specific variables in their models, so we cannot compare our data quality to theirs. However, the percent agreement and kappas we report for physiological variables and types of admission are generally considered good, while the kappa for chronic organ insufficiency is more problematic.

## Table 6.1
## Agreement statistics - training audit
## (2002-03 data only)

| Variable | Records Audited | % Exact Agreement | Kappa |
|---|---|---|---|
| Temperature, central - Low | 55 | 93 | 0.97 |
| Temperature , central - High | 55 | 91 | 0.87 |
| Mean arterial pressure (Lowest diastolic) | 55 | 80 | 0.92 |
| Mean arterial pressure (Highest diastolic) | 55 | 82 | 0.89 |
| Mean arterial pressure (Lowest systolic) | 55 | 91 | 0.95 |
| Mean arterial pressure (Highest systolic) | 55 | 91 | 0.92 |
| Heart Rate - Low | 55 | 95 | 0.98 |
| Heart Rate - High | 55 | 89 | 0.96 |
| Ventilated respiratory rate - Low | 55 | 98 | 0.96 |
| Ventilated respiratory rate - High | 55 | 82 | 0.84 |
| Non-ventilated respiratory rate - Low | 55 | 91 | 0.91 |
| Non-ventilated respiratory rate - High | 55 | 78 | 0.90 |
| $FiO_2$ <0.5 | 55 | 87 | 0.79 |
| $FiO_2$ >0.5 | 55 | 95 | 0.88 |
| Arterial pH - Low | 55 | 93 | 0.81 |
| Arterial pH - High | 55 | 95 | 0.91 |
| Serum Sodium - Low | 55 | 98 | 0.96 |
| Serum Sodium - High | 55 | 100 | 1 |
| Serum Potassium - Low | 55 | 96 | 0.97 |
| Serum Potassium - High | 55 | 98 | 0.99 |
| Serum Creatinine - Low | 55 | 93 | 0.98 |
| Serum Creatinine - High | 55 | 95 | 0.87 |
| Hematocrit - Low | 55 | 95 | 0.97 |
| Hematocrit - High | 55 | 91 | 0.83 |
| White blood cell count - Low | 55 | 91 | 0.95 |
| White blood cell count - High | 55 | 91 | 0.85 |
| Serum $HC0_3$ - Low | 55 | 98 | 0.97 |
| Serum $HC0_3$ - High | 55 | 98 | 0.98 |
| Glascow Coma Score | 55 | 63 | 0.8 |
| Chronic Organ Insufficiency | 55 | 78 | 0.29 |
| Type of Admission | 55 | 100 | 1 |

## Data Collection Process

At the individual hospitals, data were abstracted from patient medical records and were either directly entered into the CALICO data-entry software or entered into a paper-version of the data collection instrument and later transferred to the software. Periodically, the hospitals uploaded the patient information to a secure, password-protected FTP site, where it was downloaded by the CALICO project. The dataset was then read into the SAS statistical software program.

The data elements collected included all variables required to calculate mortality predictions for the SAPS II, $MPM_0$ II, APACHE II, and APACHE III models. Additional variables were collected to be evaluated for use in the generation of a new model. Some of the variables added included specific location of myocardial infarctions, a detailed description of the type of pneumonia, and an additional 12 chronic health conditions. See Table 6.2 for a listing of additional variables.

The reason for admission to the intensive care unit was abstracted by data collectors who determined a primary and, if present, a secondary diagnosis from a listing of over 1000 diagnoses adapted from a coding method developed by the Intensive Care National Audit and Research Center (ICNARC). ICNARC is a center established to undertake comparative audit and evaluative research of intensive care in the United Kingdom. The ICNARC coding method was derived empirically from textual data describing the reason for admission for 10,806 patients from the Intensive Care Society's UK APACHE II study.[1] Due to its five-tiered hierarchy, the ICNARC Coding Method allows for stepwise analysis to investigate the potential value that each level of diagnostic information adds to a prognostic model. The five tiers include a description for the type of diagnosis (surgical vs. medical), body system involved, specific site involved, process, and condition. For surgical patients, CALICO has augmented the original ICNARC coding method by adding a sixth tier to describe the surgical procedure performed. ICNARC codes were mapped to appropriate APACHE II and APACHE III diagnostic codes.

**Table 6.2**
**Listing of additional variables in the CALICO data collection**

| Physiologic Variables |
|---|
| INR |
| Platelet Count |
| Body Mass Index |
| Hemoglobin |

| Chronic Health Conditions |
|---|
| Hypertension |
| Previous myocardial infarction |
| Peripheral vascular disease |
| Congestive heart failure |
| Last known ejection fraction <40% |
| HIV+, but no AIDS related illnesses |
| Chronic obstructive pulmonary disease |
| Diabetes |
| Chronic complications of diabetes (nephropathy, neuropathy, etc.) |
| Transient ischemic attack |
| Cerebrovascular accident |
| Previous carotid artery surgery |

| Reason for Admission |
|---|
| Over 1000 ICNARC codes for reason for admission |
| Nosocomial versus community-acquired pneumonia |
| Location of myocardial infarction |
| Timing of myocardial infarction (≤24 or > 24 hrs prior to admission) |

## References:

[1] De Keizer NF, Bonsel GJ, Goldfad, Rowan KM. The added value that increasing levels of diagnostic information provide in prognostic models to estimate hospital mortality for adult intensive care patients. *Intensive Care Med.* 2000; 26: 577-584.

## Model Descriptions and Development

### Risk-Adjusted ICU Mortality Models

In order to compare the extant ICU risk-adjusted mortality models, the data elements required to calculate SAPS II, $MPM_0$ II, APACHE II, and APACHE III scores and mortality probabilities were collected. All of these models are based on multiple logistic regression equations. Programs were written for SAS to calculate SAPS II, $MPM_0$ II, APACHE II, and APACHE III scores and mortality probabilities. Descriptions of the equations as well as the coefficients calculated for each variable were obtained from the published literature for the SAPS II, $MPM_0$ II, and APACHE II models. For the APACHE III model, information was obtained from technical papers in the published literature as well as personal communications with the producers of APACHE III. When questions arose concerning the APACHE III equation, the APACHE/Cerner Corporation was contacted directly. At the time of initiation of the CALICO project, the coefficients and diagnostic categories used in the APACHE III equations were proprietary and required specific authorization by the producers of APACHE III for use in the CALICO project, but these have now been released to the public. A summary of the initial development of the ICU-risk adjusted mortality models and the variables contained in each model are found in the following section.

# SAPS II

The Simplified Acute Physiology Score (SAPS II) was developed from a large multi-center European/North American study that enrolled 14,745 patients from more than 12 countries and 137 ICUs. The underlying methodology of SAPS was to use logistic regression to help determine the variables that would be used to generate the SAPS score, to select appropriate groupings and weight assignments for each variable, and to convert the SAPS II score into a probability of hospital mortality.[1]

There were 37 initial variables collected that were chosen for clinical reasons by the coordinators of the SAPS study. For physiologic variables, the worst value in the first 24 hours after ICU admission was used. Bivariate analyses were used to identify independent variables that were associated with hospital mortality. Continuous variables that were statistically significant were plotted against vital status at hospital discharge. The LOWESS (locally weighted least squares) smoothing function was used to propose ranges for each variable.[1] Dummy variables were then assigned to each range and placed in a multiple logistic regression to determine coefficients for each of the variables. The resultant coefficients were used to assign points to each range, generally multiplying the β coefficient by ten and rounding to the nearest integer.

For an individual patient, the points assigned to each of the variables were combined to produce a SAPS II score. The subsequent SAPS II scores were used in a multiple logistic regression equation to predict the probability of inpatient mortality. The performance of the model was evaluated by Hosmer-Lemeshow tests. If variables did not improve the goodness-of-fit in the logistic regression equation, they were excluded from the study. Seventeen variables were in the final SAPS II model. (See Tables 7.1 and 7.2 for a listing of the variables included and excluded.) In examining the developmental dataset, the developers of SAPS II noted that the distribution of SAPS II scores was highly skewed. As a result, a shrinking power transformation, ln (SAPS II score +1) was incorporated into the model. The final multiple logistic equation and conversion to determine hospital mortality are noted below.

**logit = -7.7631 + 0.0737(SAPS II score) + 0.9971 [ln (SAPS II score + 1)]**

**Probability of Inpatient Mortality** $= e^{logit} / 1 + e^{logit}$

## Table 7.1
## SAPS II variables and point assignments

| Variable | Value | Pts. | Value | Pts. | Value | Pts. |
|---|---|---|---|---|---|---|
| Age | <40<br>40-59 | 0<br>7 | 60-69<br>70-74 | 12<br>15 | 75-79<br>≥ 80 | 16<br>18 |
| HR | < 40<br>40-69 | 11<br>2 | 70-119<br>120-159 | 0<br>4 | ≥ 160 | 7 |
| Systolic BP | <70<br>70-99 | 13<br>5 | 100-199<br>≥ 200 | 0<br>2 | | |
| Temp. ($^oC$) | <39$^o$ | 0 | ≥39$^o$ | 3 | | |
| $PaO_2$ (m Hg)/ $FiO_2$ | < 100 | 11 | 100-199 | 9 | ≥200 | 6 |
| Urine output (L/day) | <0.5 | 11 | 0.5-.999 | 4 | ≥1 | 0 |
| BUN (mg/dL) | <28 | 0 | 28-83 | 6 | ≥ 84 | 10 |
| WBC ($10^3/mm^3$) | < 1.0 | 12 | 1.0-19.9 | 0 | ≥ 20 | 3 |
| Serum K+ (mmol/L) | <3.0 | 3 | 3.0-4.9 | 0 | ≥5 | 3 |
| Serum Na+ (mmol/L) | < 125 | 5 | 125-144 | 0 | ≥ 145 | 1 |
| Serum $HCO_3$ (mEq/L) | <15 | 6 | 15-19 | 3 | ≥ 20 | 0 |
| Bilirubin (mg/dL) | <4.0 | 0 | 4.0-5.9 | 4 | ≥ 6 | 9 |
| Glasgow coma score | <6<br>6-8 | 26<br>13 | 9-10<br>11-13 | 7<br>5 | 14-15 | 0 |
| Chronic disease | Metastatic cancer | 9 | AIDS | 17 | Hematologic malignancy | 10 |
| Type of admission | Scheduled surgical | 0 | Unscheduled surgical | 8 | Medical | 6 |

## Table 7.2
## Non-significant variables excluded from the SAPS II model

Acute Physiology
- Respiratory Rate
- Serum glucose level
- Serum albumin level
- Serum creatinine level

Organ System Failure First Day
- Respiratory failure
- Renal failure
- Cardiovascular failure
- Hematologic failure
- Neurological failure
- Hepatic failure

Comorbid Conditions
- Cirrhosis
- Insulin-dependent diabetes
- Chronic obstructive pulmonary disease
- Heart failure
- Taking non-steroidal anti-inflammatory drugs
- Receiving chemotherapy
- Taking steroids

Previous Health State
- ABCD system (four possibilities)
- MacCabe (three classes)

** Adapted from Le Gall et. al. "A New Simplified Acute Physiology Score (SAPS II)"*

## $MPM_0II$

The Mortality Probability Model II on admission ($MPM_0$ II) was developed as an updated and revised version of the Mortality Probability Model. The goal of the developers was to construct a model that would accurately predict the mortality experience of a patient sample using the fewest variables required to discriminate and calibrate well.[2] Only variables that had clear definitions, could be easily obtained, and could be reliably collected were included in the final model. The model did not require the data collectors to obtain a single, primary reason for admission. All variables were collected at the time of ICU admission.

Bivariate analyses were carried out with each of the prospective variables to test for associations with hospital mortality. Chi-square tests were used to assess for associations with hospital mortality for categorical variables, while the Student's *t* test and Wilcoxon's Rank Sum tests were used for continuous variables. Variables were eligible to be included in the multiple logistic regression model if they were significantly associated with hospital mortality at a P value of <0.1 and constituted at least 2% of the sample population. Variables were then placed in a multivariate model and eliminated if they were not significant at a P value ≤ 0.05. Calibration of the multivariate model was assessed using the Hosmer-Lemeshow goodness-of-fit test and discrimination by the area under the receiver operating characteristic curve. In an effort to reduce the number of variables in the model, variables whose exclusion improved calibration while not significantly impacting discrimination were considered for omission from the model. Interactions between the variables were also assessed. For an interaction term to be included in the final model, it needed to be significant at a P value of ≤ 0.05, be present in at least 1% of the sample population, and be clinically plausible. No interaction terms met these criteria. The final model included 15 variables. The definitions of the variables and their coefficients included in the model are listed in Tables 7.3 and 7.4.

### Table 7.3
### Variables in the $MPM_0$ II model and their estimated coefficients

| Variable | Coefficient (β) | SE |
|---|---|---|
| Constant | -5.46836 | n/a |
| Physiology | | |
| Coma or deep stupor | 1.48592 | (0.079) |
| Heart Rate ≥ 150 beats/min | 0.45603 | (0.145) |
| Systolic blood pressure ≤ 90 mmHg | 1.06127 | (0.079) |
| Chronic Diagnoses | | |
| Chronic renal insufficiency | 0.91906 | (0.105) |
| Cirrhosis | 1.13681 | (0.126) |
| Metastatic neoplasm | 1.19979 | (0.098) |
| Acute Diagnoses | | |
| Acute renal failure | 1.48210 | (0.089) |
| Cardiac dysrhythmia | 0.28095 | (0.068) |
| | | (0.089) |
| Cerebrovascular accident | 0.21338 | |
| GI bleeding | 0.39653 | (0.094) |
| Intracranial mass effect | 0.86533 | (0.088) |
| Other | | |
| Age (10 year odds ratio) | 0.03057 | (0.002) |
| CPR prior to admission | 0.56995 | (0.112) |
| Mechanical ventilation | 0.79105 | (0.056) |
| Non-elective Surgery | 1.19098 | (0.074) |

** Adapted from Lemeshow et. al. "Mortality Probability Models (MPM II) "*

# Table 7.4.
# $MPM_0II$ variables and definitions

## Physiology

**Coma or deep stupor at ICU admission, not due to drug overdose**

- For patients taking a paralyzing muscle relaxant, awakening from anesthesia, or heavily sedated, use your best judgment of the level of consciousness prior to sedation
  - Coma - No response to any stimulation, no twitching, no movement in extremities, no response to pain or command; generally corresponds to a Glasgow Coma Scale (GCS) of 3
  - Deep stupor - Exhibits decorticate or decerebrate posturing; posturing is spontaneous or in response to stimulation or deep pain, not in response to commands; generally corresponds to a GCS of 4 or 5

**Heart rate at ICU admission**

- Record whether the heart rate was noted to be ≥ 150 beats/min within 1hr before or after ICU admission

**Systolic blood pressure at ICU admission**

- Record whether the systolic blood pressure was noted to be 90 mm Hg within 1 h before or after ICU admission

## Chronic Diagnoses

**Chronic renal compromise or insufficiency**

- Should have evidence of an elevation of serum creatinine >176.8 µmol/L (2.0 mg/dL) and be documented as chronic in the medical history; if patient *has* acute diagnosis on chronic renal failure, then only record as yes under acute renal failure

**Cirrhosis**

- Yes if history of heavy alcohol use with portal hypertension and varices, other causes with evidence of portal hypertension and varices, or biopsy confirmation

**Metastatic malignant neoplasm**

- For example, stage IV cancer; excludes regional nodes; mark yes if metastases obvious by clinical assessment or confirmed by a pathology report; mark no if it is not obvious and has to be confirmed by a pathology report that is not available at the time of ICU admission; acute hematologic malignancies are considered in this category; chronic leukemia is considered only if there are findings attributable to the disease or the patient is under active treatment for the leukemia; findings include sepsis, anemia, stroke caused by clumping of white blood cells, tumor lysis syndrome (increased uric acid as the result of chemotherapy), pulmonary edema including lymphangiectatic form of ARDS

## Acute Diagnoses

**Acute renal failure**

- Acute tubular necrosis, or acute diagnosis of chronic renal failure; prerenal failure is not included

**Cardiac dysrhythmias**

- Cardiac arrhythmia, paroxysmal tachycardia, fibrillation with rapid ventricular response, second- or third-degree heart block; do not include chronic and stable arrhythmias

**Cerebrovascular incident**

- Cerebral embolism, occlusion, CVA, stroke, brain-stem infarction, cerebrovascular arteriovenous malformation (acute stroke or cerebrovascular hemorrhage, not chronic arteriovenous malformation)

**Gastrointestinal bleeding**

- Hematemesis, melena; a perforated ulcer does not necessarily indicate Gl bleeding—may be identified by obvious coffee grounds in the NG tube; a drop in hemoglobin by itself is not sufficient evidence of acute GI bleeding

**Intracranial mass effect**

- Intracranial mass (abscess, tumor, hemorrhage, subdural) as identified by CT or other scan associated with any of the following: (I) midline shift, (2) obliteration or distortion of cerebral ventricles, (3) gross hemorrhage in cerebral ventricles or subarachnoid space, (4) visible mass >4cm, or (5) any mass that enhances with contrast media; if the mass effect is known within 1 h of ICU admission, it can be indicated as yes; CT scanning is not mandated and is only indicated for patients with major neurological insult

## Other

**Age**

- Patient's age at last birthday

**CPR within 24 h prior to ICU admission**

- CPR includes chest compression, defibrillation, or cardiac massage; record as yes regardless of where the CPR was administered

**Mechanical ventilation**

- Patient is using a ventilator at the time of ICU admission or immediately thereafter

**Medical or unscheduled surgery admission**

- This question should be answered yes except for (1) elective-surgery (i.e. surgery scheduled at least 24 h in advance) patients and (2) preoperative Swan-Ganz insertion if elective-surgery patient

***ICU*** indicates intensive care unit; ***ARDS***, adult respiratory distress syndrome; ***ATN***, acute tubular necrosis; ***CVA***, cerebrovascular accident; ***GI***, gastrointestinal; ***NG***, nasogastnc; ***CT***, computed tomography; and ***CPR***, cardiopulmonary resuscitation.

** Adapted from Lemeshow et. al. "Mortality Probability Models (MPM II) "*

## The APACHE II Model

The APACHE II model is based on an earlier prototype system, APACHE (Acute Physiology and Chronic Health Evaluation). The fundamental basis of the APACHE model is that the severity of acute illness can be quantified by the degree of abnormality in multiple physiologic variables.[3] In the original APACHE model, a variation of the nominal group process was used to choose and weight physiologic variables.[4] Thirty-four variables were chosen and each given a weight of 0-4 depending on the degree of physiologic derangement. In the APACHE II model, the number of physiologic variables was reduced to 12.[5] Infrequently collected variables such as lactic acid and serum osmolarity were excluded, as well as variables that were similar markers of disease, e.g. BUN and creatinine. For each deleted measurement, a multivariate comparison of the original APACHE system with each proposed change was evaluated to assess the impact on the statistical precision of the model. Ultimately, the fewest physiologic variables that would reflect physiologic derangement for all organ systems, while maintaining precision, were retained.[5] Of these, Glasgow Coma Score was the only measure of neurologic function. Since coma has been shown to have significant impact on mortality, it received greater weight than the other variables.[6] Additionally, the loss of renal function is known to be a strong indicator of poor prognosis, so serum creatinine was given double weight in patients with acute renal failure.[7] If multiple values for a given variable exist in the first 24 hours after ICU admission, the value with the worst derangement is used. (See Table 7.5 for listings of variables and weights.)

The APACHE II score also includes markers of diminished physiologic reserve. Both age and severe chronic disease reduce the probability of survival during an acute illness. From the original APACHE study, it was noted that patients who were non-operative or who had emergency surgery were at greater risk of death secondary to their prior organ system insufficiency than elective surgical patients. The hypothesis was that patients with the most severe chronic health conditions may not be considered candidates for elective surgery.[5] As a result, emergency surgery patients and non-operative patients are given a higher weight for severe chronic organ dysfunction than elective surgical patients. Tables 7.6 and 7.7 list the definitions of chronic severe organ dysfunction as well as the weights assigned to severe chronic organ dysfunction and age.

The weights assigned to the physiology variable derangements were combined to produce the Acute Physiology Score (APS). The APS was then added to the weights for age and chronic health to derive the APACHE II score. Using multiple logistic regression, coefficients were derived for the probability of death for 50 diagnostic categories (Table 7.8), the APS, and post-emergency surgery patients. This resulted in an equation that could be used to calculate the probability of death for an individual patient.

R = the risk of hospital death

$$\ln(R/1-R) = -3.517 + (\text{APACHE II Score} * 0.146) + (0.603, \text{ only if post emergency surgery}) + (\text{Diagnostic category coefficient})$$

## Table 7.5
## The APACHE II severity of disease classification system

| Physiologic Variable | High Abnormal Range | | | | | Low Abnormal Range | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **+4** | **+3** | **+2** | **+1** | **0** | **+1** | **+2** | **+3** | **+4** |
| Temperature – rectal (ºC) | ≥ 41º | 39º -40.9º | | 38.5º-38.9º | 36º-38.4º | 34º-35.9º | 32º-33.9º | 30º-31.9º | ≤ 29.9º |
| Mean Arterial Pressure – mm Hg | ≥ 160 | 130-159 | 110-129 | | 70-109 | | 50-69 | | ≤ 49 |
| Heart Rate | ≥ 180 | 140-179 | 110-139 | | 70-109 | | 55-69 | 40-54 | ≤ 39 |
| Respiratory Rate – ventilated or non-ventilated | ≥ 50 | 35-49 | | 25-34 | 12-24 | 10-11 | 6-9 | | ≤ 5 |
| Oxygenation<br>A. if $FiO_2 \geq 0.5$ , record $A\text{-}aDO_2$ | ≥ 500 | 350-499 | 200-349 | | <200 | | | | |
| B. if $FiO_2 < 0.5$ , record only $PaO_2$ | | | | | > 70 | 61-70 | | 55-60 | <55 |
| Arterial pH | ≥ 7.7 | 7.6-7.69 | | 7.5-7.59 | 7.33-7.49 | | 7.25-7.32 | 7.15-7.24 | <7.15 |
| Serum Sodium – mMol/L | ≥ 180 | 160-179 | 155-159 | 150-154 | 130-149 | | 120-129 | 111-119 | ≤ 110 |
| Serum Potassium – mMol/L | ≥ 7 | 6-6.9 | | 5.5-5.9 | 3.5-5.4 | 3-3.4 | 2.5-2.9 | | <2.5 |
| Serum creatinine – mg/100 cc<br>(Double pts for Acute Renal Failure) | ≥ 3.5 | 2-3.4 | 1.5-1.9 | | 0.6-1.4 | | <0.6 | | |
| Hematocrit (g%) | ≥ 60 | | 50-59.9 | 46-49.9 | 30-45.9 | | 20-29.9 | | <20 |
| White Blood Count total/$mm^3$ (x $10^3$) | ≥ 40 | | 20-39.9 | 15-19.9 | 3-14.9 | | 1-2.9 | | <1 |
| Glasgow Coma Score – Score= 15 minus actual GCS | | | | | | | | | |
| | | | | | | | | | |
| Serum $HCO_3$ (venous –mMol/l)<br>Use if No ABGs | ≥ 52 | 41-51.9 | | 32-40.9 | 22-31.9 | | 18-21.9 | 15-17.9 | <15 |

** Adapted from Knaus et. al. "Apache II: A severity of disease classification system"*

## Table 7.6
## APACHE II definitions of chronic severe organ insufficiency

| Organ insufficiency or immuno-compromised state must have been evident prior to hospital admission and conform to the following criteria: | |
|---|---|
| Liver | Biopsy proven cirrhosis and documented portal hypertension; episodes of past upper GI bleeding attributed to portal hypertension; or prior episodes of hepatic failure / encephalopathy / coma |
| Cardiac | New York Heart Association Class IV |
| Respiratory | Chronic restrictive, obstructive, or vascular disease resulting in severe exercise restriction, i.e. unable to climb stairs or perform household duties; or documented chronic hypoxia, hypercapnia, secondary polycythemia, severe pulmonary hypertension (>40mmHg), or respirator dependency |
| Renal | Receiving chronic dialysis |
| Immune | The patient has received therapy that suppresses resistance to infection, e.g.; immuno-suppression, chemotherapy, radiation, long term or recent high dose steroids, or has a disease that is sufficiently advanced to suppress resistance to infection, e.g. leukemia, lymphoma, AIDS |

** Adapted from Knaus et. al. "Apache II: A severity of disease classification system"*

## Table 7.7
## APACHE II weights for age and chronic health

| Age (yrs) | Points |
|---|---|
| ≤ 44 | 0 |
| 45-54 | 2 |
| 55-64 | 3 |
| 65-74 | 5 |
| ≥ 75 | 6 |

| **If a patient has a history of severe organ system insufficiency or is immnuno-compromised** | |
|---|---|
| Non-operative or emergency postoperative patients | 5 |
| Elective postoperative patients | 2 |

# Table 7.8
# APACHE II diagnoses and coefficients

## NONOPERATIVE PATIENTS

| | |
|---|---|
| *Respiratory failure or insufficiency from:* | |
| Asthma/allergy | -2.108 |
| COPD | -0.367 |
| Pulmonary edema (non-cardiogenic) | -0.251 |
| Postrespiratory arrest | -0.168 |
| Aspiration/poisoning/toxic | -0.142 |
| Pulmonary embolus | -0.128 |
| Infection | 0 |
| Neoplasm | 0.891 |
| *Cardiovascular failure or insufficiency from:* | |
| Hypertension | -1.798 |
| Rhythm disturbance | -1.368 |
| Congestive heart failure | -0.424 |
| Hemorrhagic shock/hypovolemia | 0.493 |
| Coronary artery disease | -0.191 |
| Sepsis | 0.113 |
| Postcardiac arrest | 0.393 |
| Cardiogenic shock | -0.259 |
| Dissecting thoracic/abdominal aneurysm | 0.731 |
| *Trauma:* | |
| Multiple trauma | -1.228 |
| Head trauma | -0.517 |
| *Neurologic:* | |
| Seizure disorder | -0.584 |
| ICH/SDH/SAH | 0.723 |
| *Other:* | |
| Drug overdose | -3.353 |
| Diabetic ketoacidosis | -1.507 |
| GI bleeding | 0.334 |
| *If not in one of the specific groups above, then use major organ system:* | |
| Metabolic/renal/hematologic | -0.885 |
| Respiratory | -0.890 |
| Neurologic | -0.759 |
| Cardiovascular | 0.470 |
| Gastrointestinal | 0.501 |

## POSTOPERATIVE PATIENTS

| | |
|---|---|
| Multiple trauma | -1.684 |
| Admission due to chronic CV disease | -1.376 |
| Peripheral vascular surgery | -1.315 |
| Heart valve surgery | -1.261 |
| Craniotomy for neoplasm | -1.245 |
| Renal surgery for neoplasm | -1.204 |
| Renal transplant | -1.042 |
| Head trauma | -0.955 |
| Thoracic surgery for neoplasm | -0.802 |
| Craniotomy for ICH/SDH/SAH | -0.788 |
| Laminectomy and other spinal cord surgery | -0.699 |
| Hemmorhagic shock | -0.682 |
| GI bleeding | -0.617 |
| GI surgery for neoplasm | -0.248 |
| Respiratory insufficiency after surgery | -0.140 |
| GI perforation/obstruction | 0.060 |
| Sepsis | 0.113 |
| Postcardiac arrest | 0.393 |
| Postrespiratory arrest | -0.168 |
| *If not in one of the specific groups above, then use major organ system:* | |
| Metabolic/renal/hematologic | -0.196 |
| Respiratory | -0.610 |
| Neurologic | -1.150 |
| Cardiovascular | -0.797 |
| Gastrointestinal | -0.613 |

** Adapted from Knaus et. al. "Apache II: A severity of disease classification system"*

## The APACHE III Model

The APACHE III model was developed in an attempt to improve on the APACHE II model. This was accomplished by re-evaluating the selection and weighting of physiologic variables, examining how differences in patient admission criteria and timing of admission to ICUs related to outcome variations among hospitals, expanding the size and representativeness of the database upon which the model was developed, and examining issues regarding the selection of patients and the timing of scoring.[8]

To estimate weights for acute physiology, chronic health conditions, and age, multivariate logistic regression analyses were used to determine the relationship between hospital mortality and the variable being assessed. In evaluating the optimal time to obtain the physiologic variables, individuals were given APACHE III scores by using the first value for the physiologic variables during the first hour of admission, the worst value over the remaining 23 hours of the first day of ICU admission, and the worst value over the initial 24 hours of ICU admission. The absolute differences between the scores were not statistically significant. Since the proportion of missing values was lowest in the worst value over the initial 24 hours and the other groups offered no changes in the explanatory power, the developers decided to use the worst value over the initial 24 hours in the APACHE III model.

In assigning weight to the physiologic variables, variables were first placed into clinically appropriate ranges and, using multiple logistic regression, coefficients for those groups were determined. The results were compared with basic clinical and physiologic principles and adjusted to reflect those principles. Additionally, cubic spline analysis was tried as an approach to assign varying weights to the variables; however, it did not substantially increase the total explanatory power and was not used in the final model. The final approach to determining the weights was primarily empiric. Derived weights from a portion of the database were adjusted to comply with clinical and physiologic principles, then validated on a separate portion of the database.

Disease-specific weighting of the variables was also examined. For congestive heart failure, a disease category with a high prevalence and mortality rate in the dataset, weights were re-estimated for certain key physiologic variables (i.e. blood pressure). The new weights however did not improve the explanatory power of the model. An additional variable was added to APACHE III to help account for the time spent in the prior location prior to admission and its effect on mortality. The developers found that adding a variable for the time spent in the emergency department again did not increase the explanatory power; however, the time spent in another inpatient location did affect inpatient mortality so a variable was added for length of stay for patients coming from other inpatient care areas.

A list of 212 disease categories was developed to determine the reason for ICU admission. Multiple logistic regressions was used to determine coefficients for these categories and their stability was assessed with regard to the weights derived and used with APACHE II and the clinical experience of the developers of the model. The criteria used for assessing stability were

the homogeneity, the cell size, and the impact of the disease on short term outcome. As a result, the initial 212 categories were reduced to 78 categories with individual coefficients.

Ultimately, the final APACHE III score is determined by adding the weights for the individual physiologic variables to determine the acute physiology score (APS). The APS is then placed in an equation that includes variables and coefficients for age, chronic health conditions, pre-ICU length of stay, location admitted from, reason for ICU admission, and whether the patient had emergency surgery or not to determine the predicted risk of mortality. A complete listing of the APACHE III reason for admission codes and coefficients, coefficients for other variables, and weights assigned to the physiologic variables can be found in the public information section of the APACHE III / Cerner Corporation Webpage at www.apache-web.com.

**References:**


[1] Le Gall J, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA.* Dec 22-29 1993; 270(24):2957-2963.

[2] Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA.* 1993; 270:2478-2486.

[3] Knaus WA, Wagner DP, Draper EA. Relationship between acute physiologic derangement and risk of death. *J Chron Dis.* 1985; 38: 295-300.

[4] Knaus WA, Zimmerman JE, Wagner DP. APACHE – acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med.* 1981; 9: 591-597.

[5] Knaus WA, Draper EA, Wagner DP, Zimmerman J. APACHE II: A severity of disease classification system. *Crit Care Med.* 1985; 13:818-829.

[6] Teres D, Brown RB, Lemeshow S. Predicting mortality of intensive care patients: The importance of coma. *Crit Care Med.* 1982; 10:86-95.

[7] Sweet SJ, Glenney CU, Fitzgibbons JP. Synergestic effect of acute renal failure and respiratory failure in the surgical intensive care unit. *Am J Surg.* 1981; 141: 492-496.

[8] Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest.* 1991; 100:1619-1636.15.

## Customization of the Models

### Re-estimation of the coefficients in the models

In prior studies in which the models were applied to populations different from the ones on which they were developed, each model has maintained adequate discrimination but has shown poor calibration in the new population. To address calibration of these models in the Californian population, CALICO customized the models to the experience of Californian patients in 2002-03 by re-estimating the coefficients used in the models (see Table 8.1). The methods used were similar to prior attempts to customize these models to new populations.[1,2,3,4,5,6,7]

For the $MPM_0$ II model, the risk equation with its original 15 variables was re-estimated, allowing new values for each of the coefficients. The SAPS II risk equation includes a coefficient preceding the SAPS score (which consists of multiple physiologic variables, chronic health conditions, and type of admission) and a coefficient preceding the natural logarithm of the SAPS score. These two coefficients were re-estimated but the internal SAPS score, including weights assigned to various physiologic variables, was not altered. Previous attempts to customize the models have kept the internal weighting of the physiology score the same and have re-estimated only the coefficients in the regression equation.

For the APACHE II model, the coefficient preceding the Acute Physiology Score (APS) and the coefficient for emergency surgery were re-estimated. Similar to the SAPS II model, the internal APS was not changed. In the APACHE II model, there is also a separate coefficient for reason for admission to the ICU. This coefficient was re-estimated in two ways. Since the size of the CALICO dataset would not permit re-estimation of coefficients for all diagnostic categories, two approaches were used. In the first approach, the values of the coefficients for the diagnostic categories were not changed from those generated by the developer; a single coefficient was estimated for the relative contribution of the diagnostic categories as a whole in predicting mortality compared to the other variables in the risk equation. In the second re-estimation model, the coefficient for each of the diagnostic categories was re-estimated. For categories that did not have enough patients to re-estimate the coefficient, the category was reclassified into a combined category (APII_REMAINDER). The model was then rerun using only those diagnostic categories that had sufficient numbers of patients to re-estimate the coefficient and with the new combined category.

The APACHE III model's coefficients were re-estimated for age, past medical history, APS, location prior to admission, and type of admission. The coefficients for the reason for admission were treated in the same way as the APACHE II model. One model was created estimating a single parameter for the diagnostic categories as a whole. The second model generated new coefficients for the diagnostic categories using the same procedure described for APACHE II. Diagnostic categories that had insufficient numbers of patients to re-estimate the coefficient were reclassified into a combined category (APIII_REMAINDER).

To re-estimate the coefficients, the dataset was divided into a developmental dataset (60% of the sample) and validation dataset (40% of the sample). The developmental dataset was generated by taking a random sample of the complete dataset and the validation dataset consisted of the

remaining individuals in the dataset. This sampling was repeated three times to generate three developmental data sets and three sets of coefficients for each variable in each model.

## TABLE 8.1
## Re-estimated models

### $MPM_0$ II

**Original Model**
Logit = -5.468 + 1.486(coma) + 0.486 (HR>150) … 1.191(non-elective surgery)

**$MPM_0$ II Re-estimated Model**
Logit = y + $\beta_1$ (coma) + $\beta_2$ (HR>150) … $\beta_{15}$ (non-elective surgery)

### SAPS II

**SAPS II Original Model**
Logit = -7.7631 + 0.0737(SAPS II score) + 0.9971 [ln (SAPS II score + 1)]

**SAPS II Re-estimated Model**
Logit = y + $\beta_1$ (SAPS II score) + $\beta_2$[ln (SAPS II score + 1)]

### APACHE II

**Original Model**
Logit = -3.517 + (APACHE II Score * 0.146)
+ (0.603, only if post emergency surgery)
+ (Diagnostic category coefficient)

**APACHE II Re-estimated Model #1**
Logit = -y + $\beta_1$ (APACHE II Score) + $\beta_2$(if post-emergency surgery)
+ $\beta_3$(Diagnostic category coefficient as specified in the original model)

**APACHE II Re-estimated Model #2 (Diagnostic coefficients re-estimated)**
Logit = -y + $\beta_1$ (APACHE II Score) + $\beta_2$ (if post -mergency surgery)
+ $\beta_3$ (Diagnostic category 1) + $\beta_4$ (Diagnostic category 2)
+ …$\beta_i$ (Diagnostic category i) + $\beta_j$ (APII_REMAINDER)

### APACHE III

**Original Model**
Logit = -6.413 + 0.342(Age Category 1) + …1.024(PMH category 1) …0.088 (APS)….
+ (Diagnostic category coefficient)

**APACHE III Re-estimated Model #1**
Logit = -y + $\beta_1$ (Age Category 1) + …$\beta$ (PMH category 1) …$\beta_{15}$ (APS)….
+ $\beta_{26}$ (Diagnostic category coefficient as specified in the original model)

**APACHE III Re-estimated Model #2 (Diagnostic coefficients re-estimated)**
Logit = -y + $\beta_1$ (Age Category 1) + …$\beta$ (PMH category 1) …$\beta_{15}$ (APS)….
+ $\beta_3$ (Diagnostic category 1) + $\beta_4$ (Diagnostic category 2)
+ …$\beta_i$ (Diagnostic category i) + $\beta_j$ (APIII_REMAINDER)

**Y= intercept ; β = coefficient ; PMH = Past Medical History**

## References:

[1] Zhu B-P, Lemeshow S, Hosmer DW, et al. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization. *Crit Care Med.* Jan 1996;24(1):57-63.

[2] Murphy-Filkins RL, Teres D, Lemeshow S, et al. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model. *Crit Care Med.* Dec 1996;24(12):1968-1973.

[3] Glance LG, Osler TM, Dick AW. Identifying quality outliers in a large, multiple-institution database by using customized versions of the Simplified Acute Physiology Score II and the Mortality Probability Model $II_0$. *Crit Care Med.* 2002;30(9):1995-2002.

[4] Markgraf R, Deutschinoff G, Pientka L, et al. Performance of the score systems Acute Physiology and Chronic Health Evaluation II and III at an interdisciplinary intensive care unit, after customization. *Crit Care Med.* Jan 2001;5(1):31-36.

[5] Rivera-Fernandez R, Vazquez-Mata G, Bravo M, et al. The Apache III prognostic system: customized mortality predictions for Spanish ICU patients. *Intensive Care Med.* 1998;24:574-581.

[6] Metnitz PGH, Valentin A, Vesely H, et al. Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. *Intensive Care Med.* 1999;25:192-197.

[7] Sirio CA, Shepardson LB, Rotondi AJ, et al. Community-wide assessment of intensive care outcomes using a physiologically based prognostic measure. *Chest.* 1999;115(3):793-801.

## Results I: Performance of the Original Models

### Goodness-of-Fit – Original Models

***Discrimination***

Discrimination was assessed by using the area under the receiver operating characteristic curve (AUC). The minimum criterion for AUC that was considered reasonable discrimination was 0.80.[1] The AUC was determined using the 4,630 patients in the CALICO dataset that could be used across all models. $MPM_0$ II, SAPS II, APACHE II, and APACHE III all showed reasonable discrimination with AUCs of 0.804, 0.871, 0.847, and 0.884 respectively. The AUC for the $MPM_0$ II model was statistically significantly lower than the other three models. For comparison, AUCs were generated using all patients that were eligible for an individual model. This yielded almost identical AUCs compared to the analysis using only patients included in all four models.

### Table 9.1
### Summary of discrimination and calibration – original models

| Model | AUC* (95% CI) | H-L† Statistic | |
|---|---|---|---|
| | | C Test | H Test |
| **$MPM_0$ II** | 0.804 (0.787-0.821) | 47.61 (p<0.0001) | 53.10 (p<0.0001) |
| **SAPS II** | 0.871 (0.856-0.885) | 219.83 (p<0.0001) | 230.67 (p<0.0001) |
| **APACHE II** | 0.847 (0.832-0.862) | 209.20 (p<0.0001) | 212.70 (p<0.0001) |
| **APACHE III** | 0.884 (0.871-0.897) | 42.11 (p<0.0001) | 42.74 (p<0.0001) |

* = Area under the receiver operating characteristic curve
† = Hosmer-Lemeshow statistic; *df* 10

***Calibration***

Calibration was assessed using Hosmer-Lemeshow (H-L) goodness-of-fit tests and calibration curves. Both Hosmer-Lemeshow C tests and H tests were performed. Analyses using the C test divide patients into deciles (i.e. equal number of patients) in ascending order of death. The range of predicted risk of mortality within each decile is determined by the patients in that decile. The H test forms 10 groups based on fixed, equal deciles of risk (i.e. 0.0-0.09%, 0.1%-0.19%, etc.) with variable numbers of patients in each group. The difference between the observed and expected mortality for each strata is summarized by the Pearson chi-square statistic. The statistics are summed over the ten deciles and are compared to a chi-square distribution. The degrees of freedom equal N-2, where N= the number of groups, when used on an estimation dataset. However, when used on an application dataset, one in which the coefficients used are not recalculated using the dataset being analyzed, typically the degrees of freedom are the same as the number of groups (10 degrees of freedom).[1]

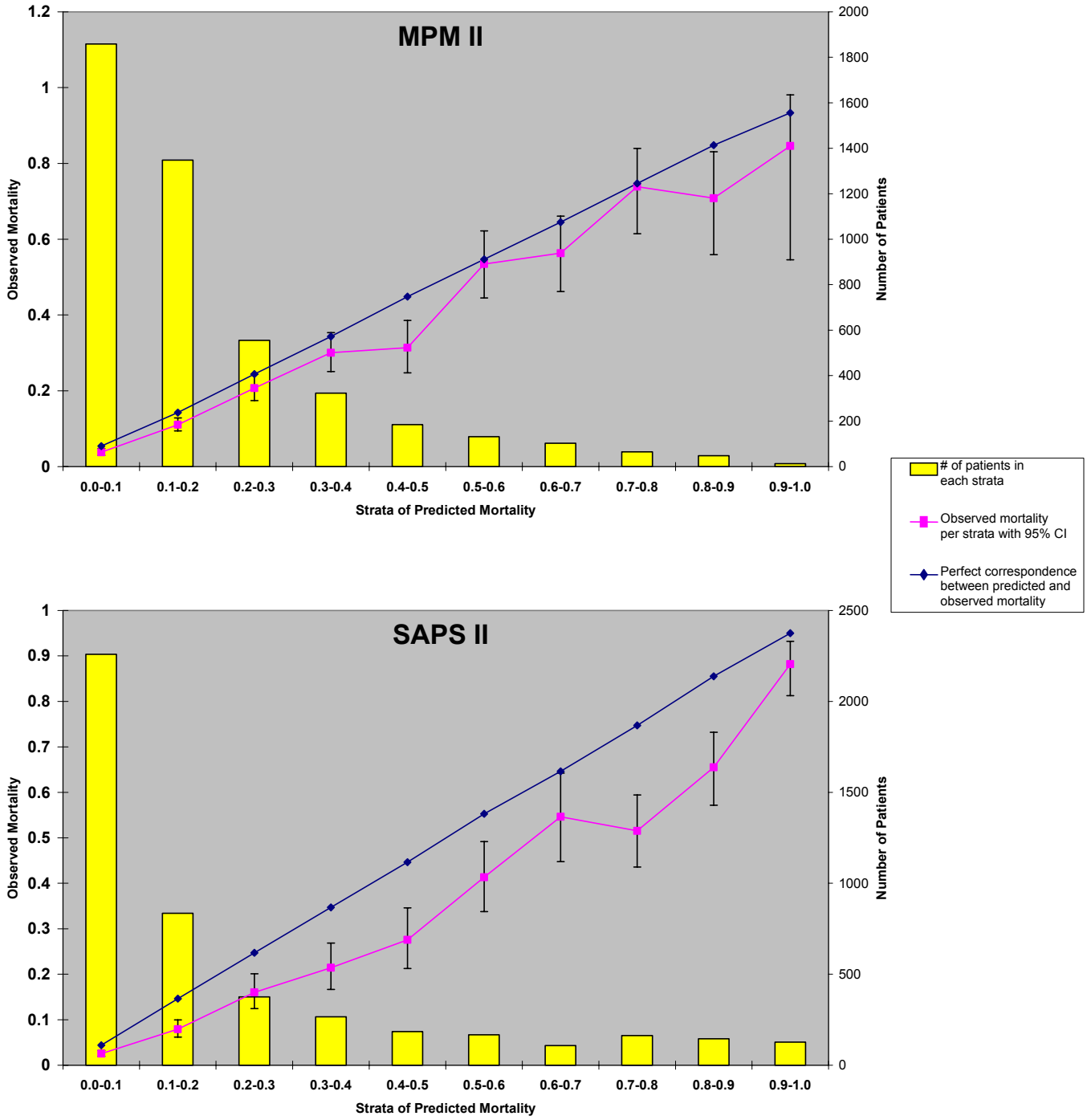
All of the models tested had calibration limitations when using their original coefficients. A complete summary of the Hosmer-Lemeshow goodness-of-fit tests can be found in Appendix Tables A.1 and A.2. The four models all had significant p-values for their Hosmer-Lemeshow

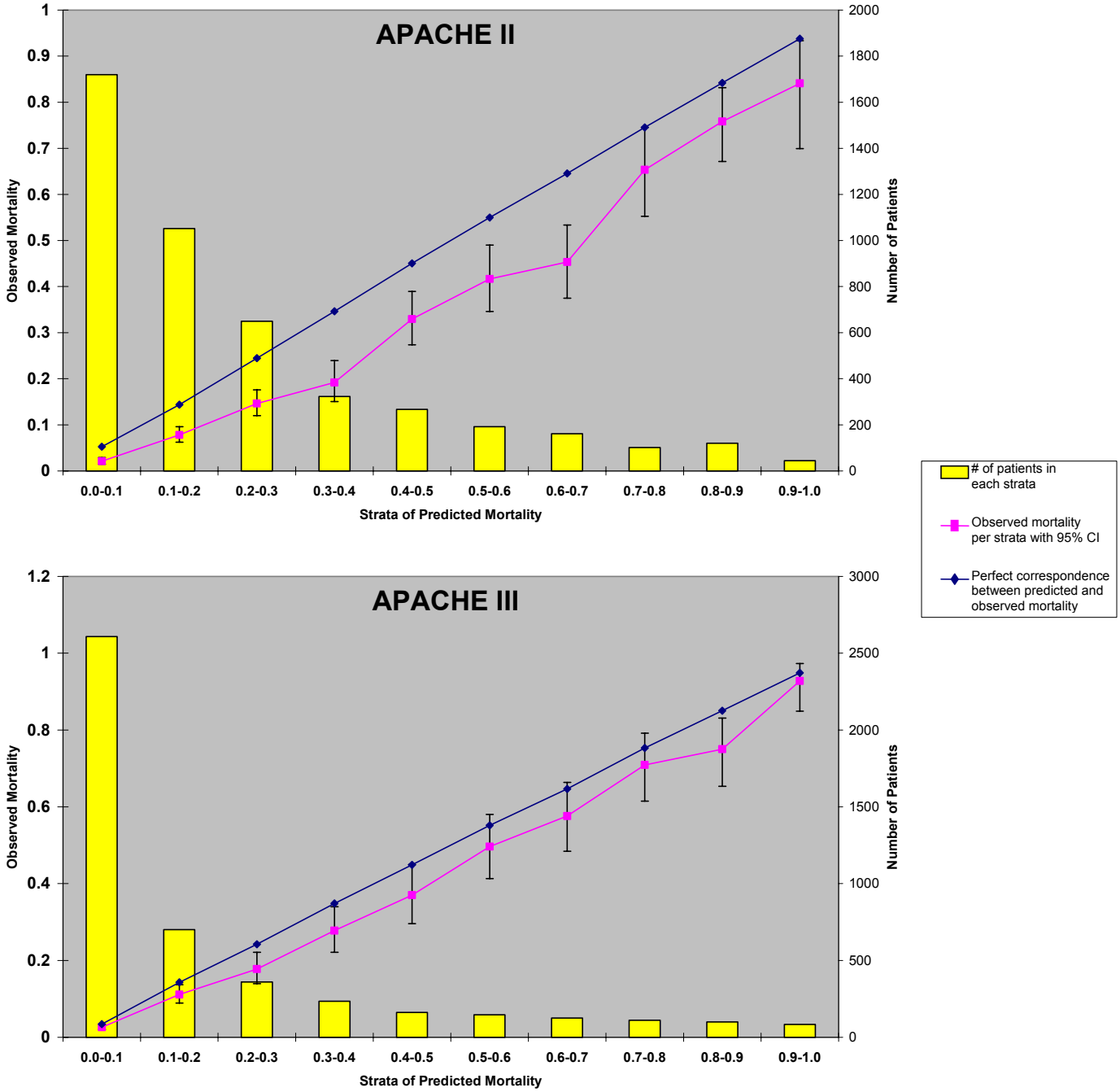
statistics, indicating a significant difference between the observed and predicted mortality, suggesting poor calibration. Although exact comparisons cannot be made among the Hosmer-Lemeshow statistics, $MPM_0$ II and APACHE III had the lower values for the Hosmer-Lemeshow statistics compared to APACHE II and SAPS II.

The calibration curves for the respective models are shown in Figure 9.1. In the calibration curves, patients are divided into strata based upon their predicted risk of mortality (0-10%, 11-20%, etc). The actual mortality rate of patients in each stratum (the number of deaths divided by the number of patients in each stratum) is plotted and compared to the line where observed mortality = expected mortality. Models showing good calibration should approximate this line. All the models over-predicted death across the ten strata of mortality risk. Comparing the models in Figure 9.1, $MPM_0$ II and APACHE III more closely approximate the observed = expected line (O=E line) than the other models. All the models are also reasonably close to the O=E line in the two lowest deciles of risk. However, there is significant variation among the other deciles of mortality risk illustrating poor uniformity of fit across deciles of predicted mortality.

Overall, $MPM_0$ II had the worst discrimination of the four models. While the AUC for the $MPM_0$ II model met the minimum criterion for adequate discrimination, it was statistically lower than other models. The models with the best calibration were the $MPM_0$ II and APACHE III models. Although these models showed better calibration than the SAPS II and APACHE II model, calibration was still poor. Before models can be used to compare the performance of ICUs, calibration should be improved. If a model over-predicts the risk of death in certain risk categories, as these models do, and a particular ICU admits disproportionately more cases in these categories, those ICUs would be incorrectly classified as having “less than predicted” deaths.

**Figure 9-1**
**Calibration curves - original models**






## References:

[1] Mourouga P, Goldfrad C, Rowan KM. Does it fit? Is it good? Assessment of scoring systems. *Current Opinion in Critical Care.* 2000;6:176–180.

## Results II: Performance of the Customized Models

***Goodness-of-Fit – Customized, Re-estimated Models***

The performance of the re-estimated models was assessed by evaluating their discrimination and calibration. Their performance was evaluated in the developmental and validation dataset across all three splits of the data. There was little difference in the AUC (discrimination) among the three different splits of the data. The re-estimated models resulted in significant improvement in the Hosmer-Lemeshow statistics (calibration). The original models all showed poor fit (P value <0.05) for both the C and H tests. In the re-estimated models, the C test indicated good fit (P value >0.05) for all models in all splits of the data. For the H test, there was also good fit across models and splits except for the validation sample of the second split for $MPM_0$ II, SAPS, and APACHE III. The results of the calibration and discrimination are summarized for split #1 in Table 10.1.

### Table 10.1
### Summary of discrimination and calibration - re-estimated models (Split #1)

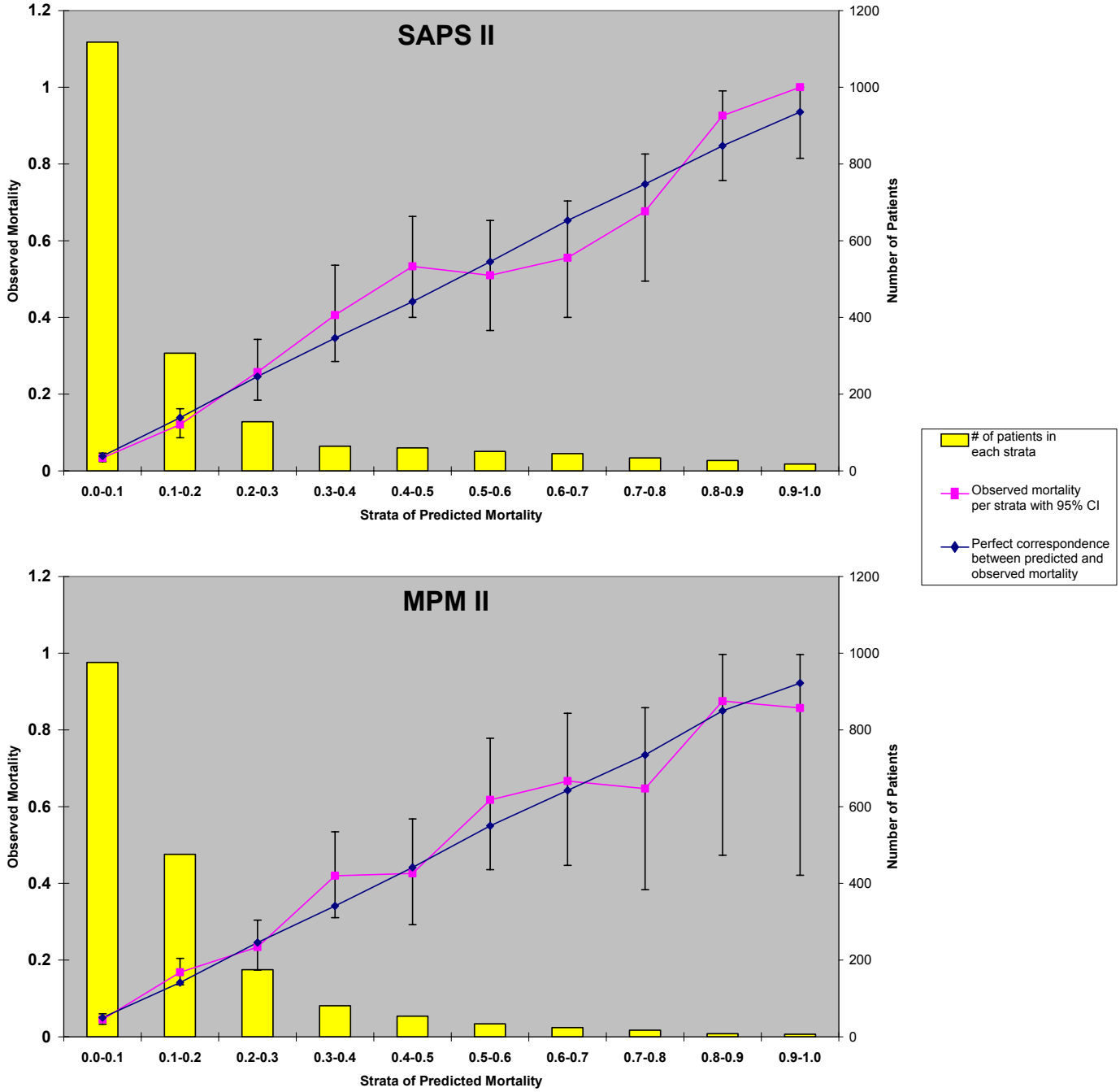
| Model | Dataset | AUC† (95% CI) | H-L‡ Statistic C Test | H-L‡ Statistic H Test |
|---|---|---|---|---|
| **MPM II** | | | | |
| **Original** | Development | 0.807 (0.784 – 0.829) | 31.9 (P<0.001) | 41.0(P<0.001) |
| | Validation | 0.800 (0.773 – 0.827) | 18.4 (P<0.05) | 18.2 (P=0.05) |
| **Re-estimated Model** | Development | 0.815 (0.793 – 0.837) | 11.3 (P=0.3) | 8.1 (P=0.6) |
| | Validation | 0.806 (0.779 – 0.834) | 6.1 (P=0.8) | 7.6 (P=0.7) |
| **SAPS II** | | | | |
| **Original** | Development | 0.866 (0.847 – 0.886) | 129.7 (P<0.001) | 138.6 (P<0.001) |
| | Validation | 0.877 (0.855 – 0.899) | 95.6 (P<0.001) | 105.7 (P<0.001) |
| **Re-estimated Model** | Development | 0.866 (0.847 – 0.886) | 8.2 (P=0.6) | 8.7 (P=0.6) |
| | Validation | 0.877 (0.855 – 0.899) | 8.5 (P=0.6) | 10.3 (P=0.5) |
| **APACHE II** | | | | |
| **Original** | Development | 0.844 (0.824 – 0.864) | 129.8 (P<0.001) | 129.3 (P<0.001) |
| | Validation | 0.852 (0.828 – 0.876) | 83.2 (P<0.001) | 88.5 (P<0.001) |
| **Re-estimated Model** | Development | 0.863 (0.844 – 0.881) | 11.2 (P=0.3) | 6.0 (P=0.8) |
| | Validation | 0.857 (0.833 – 0.881) | 3.5 (P=0.97) | 7.4 (P=0.7) |
| **Re-estimated Model with diagnostic coefficients re-estimated** | Development | 0.857 (0.838 – 0.875) | 16.6 (P=0.08) | 13.4 (P=0.2) |
| | Validation | 0.859 (0.835 – 0.882) | 7.3 (P=0.7) | 4.2 (P=0.94) |
| **APACHE III** | | | | |
| **Original** | Development | 0.884 (0.867 – 0.902) | 28.5 (P<0.001) | 32.2 (P<0.001) |
| | Validation | 0.884 (0.863 – 0.876) | 21.7 (P<0.05) | 17.6 (P=0.06) |
| **Re-estimated Model** | Development | 0.889 (0.872 – 0.906) | 2.7 (P=0.98) | 6.9 (P=0.7) |
| | Validation | 0.881 (0.860 – 0.903) | 10.8 (P=0.4) | 17.3 (P=0.07) |
| **Re-estimated Model with diagnostic coefficients re-estimated** | Development | 0.889 (0.872 – 0.906) | 2.9 (P=0.98) | 4.3 (P=0.94) |
| | Validation | 0.882 (0.861 – 0.904) | 5.9 (P=0.8) | 14.0 (P=0.2) |

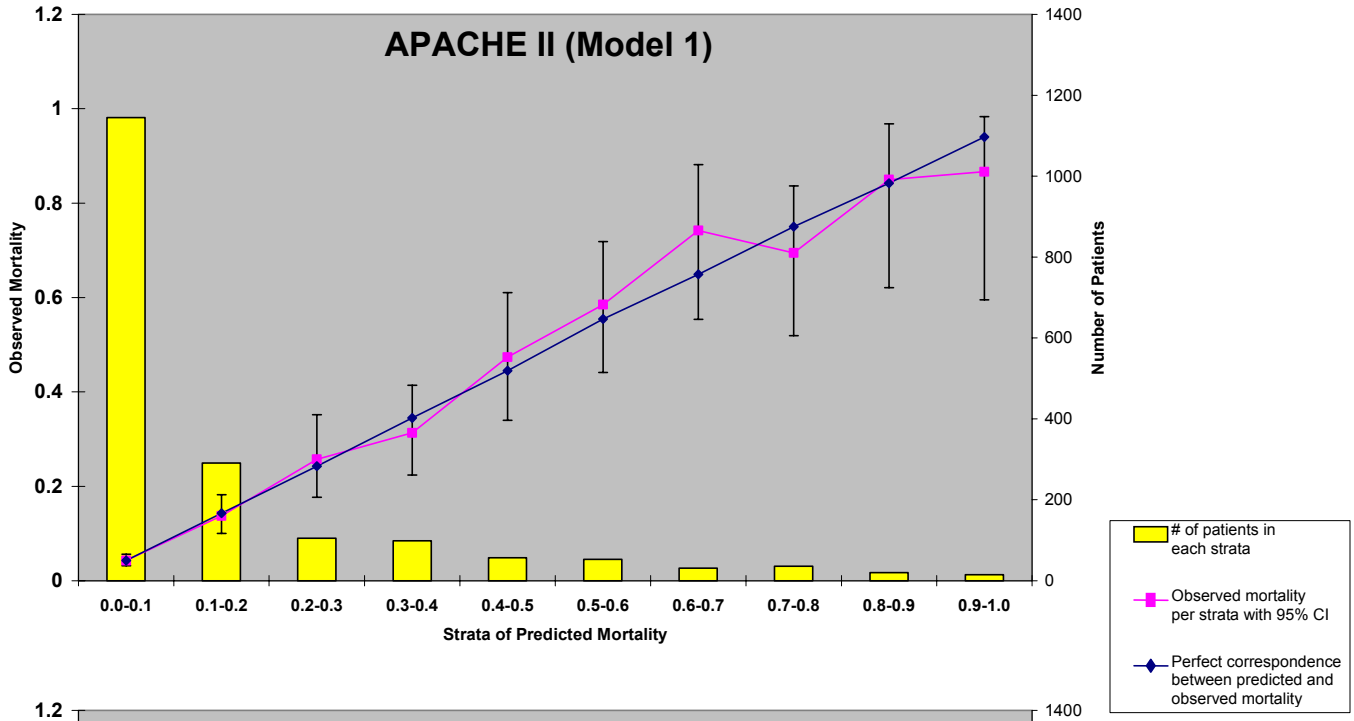
†= Area under the receiver operator curve ‡= Hosmer-Lemeshow statistic; *df* 10 for developer model; *df* 8 for re-estimated models

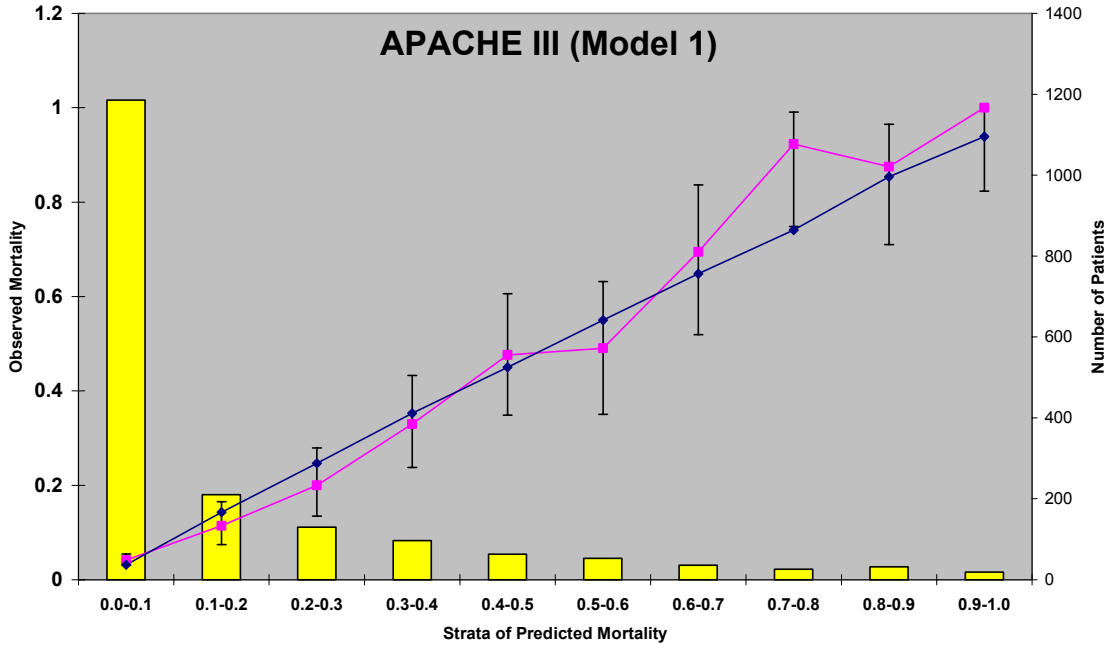
The calibration of the re-estimated models was also assessed with calibration curves (Figure 10.1). The validation dataset of the first 60/40 split of the CALICO database was used to generate the calibration curves. The curves demonstrate that with re-estimation, all of the

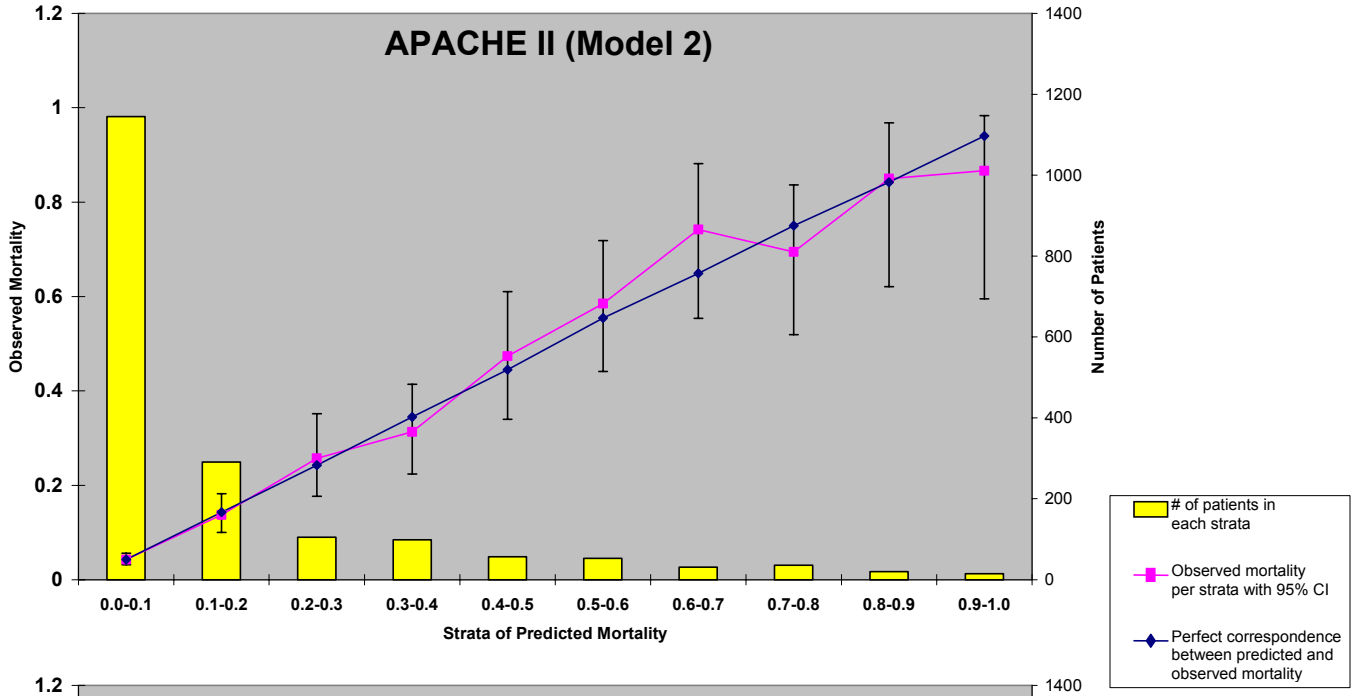
models approximate the observed deaths more closely than the original models. In the original models, the 95% confidence interval for observed mortality across the deciles of risk was often below the line for perfect correspondence, representing over-prediction of death. In general, omnibus test approaches are not as good at finding departures as more focused tests. Several data points in a row that are all well above or well below the predicted line may suggest more inaccuracy in the model than if they were spread out seemingly at random. This may imply more inaccuracy in the SAPS II and APACHE III models.

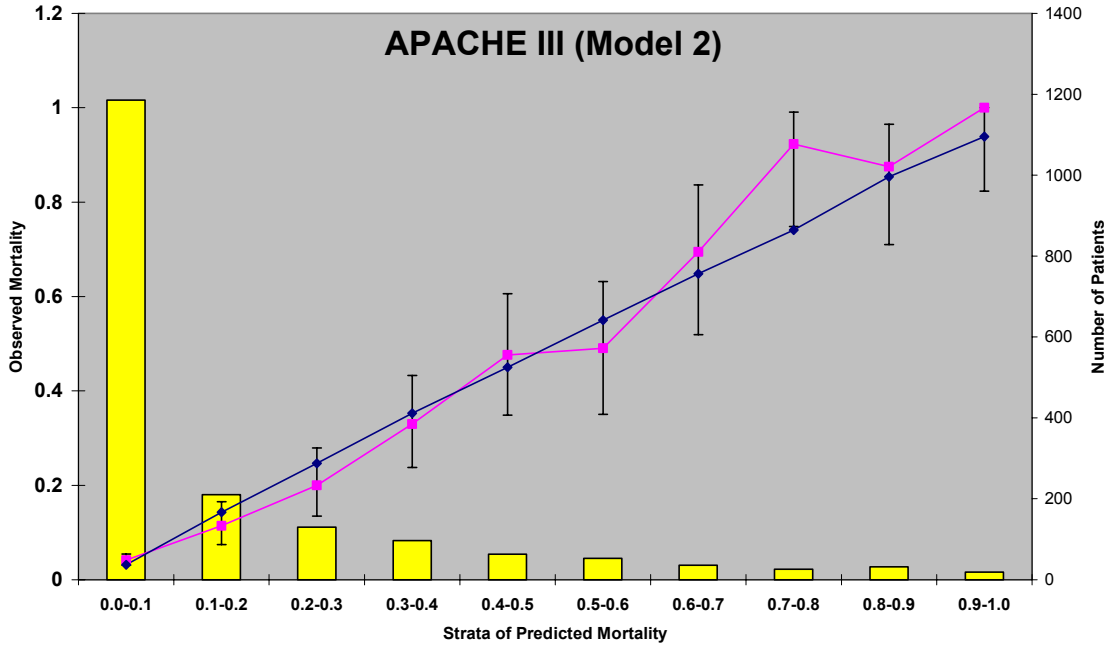
**Figure 10.1**
**Calibration curves – re-estimated models**




APACHE II (Model 1)
Observed Mortality
Number of Patients
Strata of Predicted Mortality
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1.0
# of patients in each strata
Observed mortality per strata with 95% CI
Perfect correspondence between predicted and observed mortality


APACHE III (Model 1)
Observed Mortality
Number of Patients
Strata of Predicted Mortality
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1.0


APACHE II (Model 2)
Observed Mortality
Number of Patients
Strata of Predicted Mortality
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1.0
# of patients in each strata
Observed mortality per strata with 95% CI
Perfect correspondence between predicted and observed mortality


APACHE III (Model 2)
Observed Mortality
Number of Patients
Strata of Predicted Mortality
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1.0

***Coefficients in the Customized Models***

The coefficients estimated for each split of the data are summarized in the Appendix in Tables A.3 - A.6. The re-estimated coefficients did not vary in a statistically significant way across the different splits of the data. Since the performance of the models did not vary markedly across the splits and the coefficients were similar, the entire data set was also used to generate re-estimated coefficients for the four models. Tables 10.2 - 10.5 display the re-estimated coefficients and their odds ratios for the models using the entire CALICO dataset.

## Table 10.2
## $MPM_0$ II re-estimated coefficients and odds ratios

| Variable | Original | | Re-estimated Model | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Coefficient | OR | Coefficient | | OR | 95% CI | | |
| Intercept | **-5.468** | n/a | **-6.101** | * | n/a | | | |
| Coma or deep stupor | **1.486** | 4.42 | **1.5449** | * | 4.69 | 3.44 | - | 6.39 |
| Heart Rate ≥ 150 beats/min | **0.456** | 1.58 | **1.1905** | * | 3.29 | 2.13 | - | 5.09 |
| Systolic blood pressure ≤ 90 mmHg | **1.061** | 2.89 | **0.9066** | * | 2.48 | 2.01 | - | 3.05 |
| Chronic renal insufficiency | **0.919** | 2.51 | **0.8894** | * | 2.43 | 1.71 | - | 3.47 |
| Cirrhosis | **1.137** | 3.12 | **1.0397** | * | 2.83 | 1.83 | - | 4.36 |
| Metastatic neoplasm | **1.200** | 3.32 | **1.1392** | * | 3.12 | 2.19 | - | 4.47 |
| Acute renal failure | **1.482** | 4.40 | **1.1511** | * | 3.16 | 1.93 | - | 5.19 |
| Cardiac dysrhythmia | **0.281** | 1.32 | **-0.5615** | * | 0.57 | 0.39 | - | 0.83 |
| Cerebrovascular accident | **0.213** | 1.24 | **0.553** | * | 1.74 | 1.18 | - | 2.57 |
| GI bleeding | **0.397** | 1.49 | **-0.2225** | NS | 0.80 | 0.52 | - | 1.23 |
| Intracranial mass effect | **0.865** | 2.38 | **0.5812** | * | 1.79 | 1.21 | - | 2.64 |
| CPR prior to admission | **0.570** | 1.77 | **1.4263** | * | 4.16 | 2.90 | - | 5.98 |
| Mechanical ventilation | **0.791** | 2.21 | **0.632** | * | 1.88 | 1.53 | - | 2.31 |
| Non-elective surgery | **1.191** | 3.29 | **1.118** | * | 3.06 | 2.27 | - | 4.12 |
| Age | **0.031** | 1.03 | **0.0383** | * | 1.04 | 1.03 | - | 1.05 |

*** = Coefficient was significant (P value < 0.05) in logistic regression**

## Table 10.3
## SAPS II re-estimated coefficients and odds ratios

| Variable | Original | | Re-estimated Model | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Coefficient | OR | Coefficient | | OR | 95% CI | | |
| Intercept | **-7.763** | n/a | **-8.368** | * | n/a | | | |
| SAPS II Score | **0.074** | 1.08 | **0.064** | * | 1.07 | 1.04 | - | 1.09 |
| LOG SAPS II Score | **0.997** | 2.71 | **1.095** | NS | 2.99 | 0.98 | - | 9.09 |

*** = Coefficient was significant (P value < 0.05) in the logistic regression**

# Table 10.4
# APACHE II re-estimated coefficients and odds ratios

| Variables | Original | | Re-estimated Model (1)† | | | Re-estimated Model (2)‡ | | |
|---|---|---|---|---|---|---|---|---|
| | Coefficient | OR | Coefficient | OR | 95% CI | Coefficient | OR | 95% CI |
| Intercept | **-3.517** | n/a | **-4.862** * | n/a | | **-4.665** * | n/a | |
| APACHE II score | **0.146** | 1.16 | **0.173** * | 1.19 | 1.17 – 1.20 | **0.171** * | 1.19 | 1.17 – 1.20 |
| Emergency surgery | **0.603** | 1.83 | **0.405** NS | 1.50 | 0.92 – 2.45 | **0.646** * | 1.91 | 1.07 – 3.41 |
| Diagnostic categories§ | **n/a** | n/a | **0.525** * | 1.69 | 1.45 – 1.97 | **n/a** | n/a | |
| Coronary artery disease | **-0.191** | 0.83 | - | - | - | **-0.101** NS | 0.90 | 0.61 – 1.35 |
| Metabolic/renal/hematologic non-operative | **-0.885** | 0.41 | - | - | - | **-0.686** * | 0.50 | 0.34 – 0.75 |
| Rhythm disturbance | **-1.368** | 0.25 | - | - | - | **-0.599** * | 0.55 | 0.33 – 0.91 |
| Congestive heart failure | **-0.424** | 0.65 | - | - | - | **-0.6** * | 0.55 | 0.34 – 0.88 |
| GI bleeding | **0.334** | 1.40 | - | - | - | **-0.399** NS | 0.67 | 0.40 – 1.14 |
| Neurologic non-operative | **-0.759** | 0.47 | - | - | - | **0.066** NS | 1.07 | 0.67 – 1.71 |
| ICH/SDH/SAH | **0.723** | 2.06 | - | - | - | **0.626** * | 1.87 | 1.22 – 2.88 |
| COPD | **-0.367** | 0.69 | - | - | - | **-0.334** NS | 0.72 | 0.45 – 1.14 |
| Sepsis | **0.113** | 1.12 | - | - | - | **-0.221** NS | 0.80 | 0.52 – 1.24 |
| Cardiovascular post-operative | **-0.797** | 0.45 | - | - | - | **-0.858** * | 0.42 | 0.21 – 0.85 |
| Pulmonary edema (non-cardiogenic) | **-0.251** | 0.78 | - | - | - | **-0.438** NS | 0.65 | 0.39 – 1.08 |
| Drug overdose | **-3.353** | 0.03 | - | - | - | **-2.05** * | 0.13 | 0.04 – 0.43 |
| Cardiovascular non-operative | **0.47** | 1.60 | - | - | - | **-0.458** NS | 0.63 | 0.32 – 1.27 |
| Metabolic/renal/hematologic post-operative | **-0.196** | 0.82 | - | - | - | **-2.541** * | 0.08 | 0.01 – 0.65 |
| Gastrointestinal non-operative | **0.501** | 1.65 | - | - | - | **0.077** NS | 1.08 | 0.60 – 1.95 |
| Craniotomy for neoplasm | **-1.245** | 0.29 | - | - | - | **-1.598** * | 0.20 | 0.05 – 0.87 |
| Respiratory non-operative | **-0.89** | 0.41 | - | - | - | **0.27** NS | 1.31 | 0.75 – 2.30 |
| Thoracic surgery for neoplasm | **-0.802** | 0.45 | - | - | - | **-0.751** NS | 0.47 | 0.14 – 1.56 |
| GI perforation/obstruction | **0.06** | 1.06 | - | - | - | **-0.121** NS | 0.89 | 0.41 – 1.93 |
| Hemorrhagic shock/hypovolemia | **0.493** | 1.64 | - | - | - | **-0.577** NS | 0.56 | 0.26 – 1.20 |
| Gastrointestinal post-operative | **-0.613** | 0.54 | - | - | - | **-0.552** NS | 0.58 | 0.23 – 1.43 |
| Peripheral vascular surgery | **-1.315** | 0.27 | - | - | - | **-1.098** NS | 0.33 | 0.10 – 1.14 |
| GI surgery for neoplasm | **-0.248** | 0.78 | - | - | - | **-0.754** NS | 0.47 | 0.14 – 1.57 |
| Other diagnoses‖ | **n/a** | n/a | - | - | - | **1.198** * | 3.31 | 2.10 – 5.23 |

† = Diagnostic category coefficients were NOT re-estimated
‡ = Diagnostic category coefficients were re-estimated
*** = Coefficient was significant (P value < 0.05) in logistic regression**
§ = Relative contribution of the diagnostic categories as a whole in predicting mortality compared to the other variables in the risk equation.
‖ = Diagnostic categories with insufficient patients to generate a coefficient were combined into a single category

## Table 10.5
## APACHE III re-estimated coefficients and odds ratios

| Variables | | Original | | Re-estimated Model (1)† | | | | Re-estimated Model (2)‡ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Coefficient | OR | Coefficient | | OR | 95% CI | Coefficient | | OR | 95%CI |
| Intercept: | | **-6.413** | n/a | **-5.573** | * | n/a | | **-5.523** | * | n/a | |
| Age: | 45-54 | **0.342** | 1.41 | **-0.150** | NS | 0.86 | 0.55 - 1.34 | **-0.144** | NS | 0.87 | 0.56 – 1.35 |
| | 55-59 | **0.321** | 1.38 | **0.467** | NS | 1.60 | 0.99 – 2.58 | **0.515** | * | 1.67 | 1.03 – 2.71 |
| | 60-64 | **0.613** | 1.85 | **0.493** | * | 1.64 | 1.03 – 2.61 | **0.523** | * | 1.69 | 1.05 – 2.71 |
| | 65-69 | **0.757** | 2.13 | **0.505** | * | 1.66 | 1.04 – 2.65 | **0.534** | * | 1.71 | 1.06 – 2.74 |
| | 70-74 | **1.006** | 2.74 | **0.903** | * | 2.47 | 1.63 – 3.74 | **0.935** | * | 2.55 | 1.67 – 3.88 |
| | 75-84 | **1.127** | 3.09 | **1.220** | * | 3.39 | 2.33 – 4.93 | **1.263** | * | 3.54 | 2.41 – 5.19 |
| | ≥85 | **1.495** | 4.46 | **1.328** | * | 3.77 | 2.46 – 5.79 | **1.350** | * | 3.86 | 2.50 – 5.96 |
| Past Medical History: | AIDS | **1.024** | 2.78 | **-0.336** | NS | 0.72 | 0.18 – 2.92 | **-0.344** | NS | 0.71 | 0.17 – 2.89 |
| | Hepatic failure | **1.13** | 3.11 | **0.433** | NS | 1.54 | 0.69 – 3.46 | **0.379** | NS | 1.46 | 0.65 – 3.27 |
| | Lymphoma | **1.005** | 2.73 | **1.465** | * | 4.33 | 1.77 – 10.60 | **1.459** | * | 4.30 | 1.73 – 10.67 |
| | Metastatic cancer | **0.886** | 2.43 | **0.720** | * | 2.05 | 1.24 – 3.39 | **0.727** | * | 2.07 | 1.25 – 3.43 |
| | Leukemia/multiple myeloma | **0.756** | 2.13 | **0.479** | NS | 1.62 | 0.71 – 3.69 | **0.464** | NS | 1.59 | 0.69 – 3.65 |
| | Immunosuppression | **0.321** | 1.38 | **0.327** | NS | 1.39 | 0.73 – 2.63 | **0.358** | NS | 1.43 | 0.75 – 2.72 |
| | Cirrhosis | **0.860** | 2.36 | **0.510** | NS | 1.67 | 0.90 – 3.09 | **0.481** | NS | 1.62 | 0.87 – 3.00 |
| Acute Physiology Score: | APS | **0.088** | 1.09 | **0.043** | * | 1.04 | 1.00 – 1.09 | **0.043** | * | 1.04 | 1.00 – 1.09 |
| | APS3ALL1 | **-0.309** | 0.73 | **0.250** | NS | 1.28 | 0.82 – 2.01 | **0.242** | NS | 1.27 | 0.81 – 2.00 |
| | APS3ALL2 | **0.513** | 1.67 | **-0.499** | NS | 0.61 | 0.27 – 1.37 | **-0.484** | NS | 0.62 | 0.27 – 1.39 |
| | APS3ALL3 | **-0.353** | 0.70 | **0.398** | NS | 1.49 | 0.75 – 2.96 | **0.384** | NS | 1.47 | 0.74 – 2.93 |
| | APS3ALL4 | **0.451** | 1.57 | **-0.066** | NS | 0.94 | 0.35 – 2.49 | **-0.053** | NS | 0.95 | 0.36 – 2.54 |
| Location: | Admitted to ICU from floor | **0.048** | 1.05 | **0.224** | NS | 1.25 | 0.97 – 1.62 | **0.257** | NS | 1.29 | 0.99 – 1.68 |
| | Transfer | **0.206** | 1.23 | **0.403** | NS | 1.50 | 0.80 – 2.82 | **0.374** | NS | 1.45 | 0.77 – 2.75 |
| | Admit to ICU from OR | **-0.238** | 0.79 | **-0.280** | NS | 0.76 | 0.50 – 1.13 | **-0.385** | NS | 0.68 | 0.44 – 1.05 |
| Other: | Emergency surgery | **0.079** | 1.08 | **0.416** | NS | 1.52 | 0.82 – 2.81 | **0.427** | NS | 1.53 | 0.83 – 2.84 |
| | Pre-ICU LOS (days) | **0.141** | 1.15 | **0.198** | * | 1.22 | 1.11 – 1.33 | **0.193** | * | 1.21 | 1.11 – 1.33 |
| | Excess hospital LOS | **0.069** | 1.07 | **0.099** | * | 1.10 | 1.03 – 1.19 | **0.097** | * | 1.10 | 1.02 - 1.19 |
| Diagnosis: | Diagnostic categories§ | **n/a** | n/a | **0.855** | * | 2.35 | 1.95 – 2.84 | **n/a** | | n/a | |
| | Acute myocardial infarction | **0.678** | 1.97 | - | | - | - | **0.326** | NS | 1.39 | 0.92 – 2.09 |
| | Respiratory medical other | **0.220** | 1.25 | - | | - | - | **-0.107** | NS | 0.90 | 0.59 – 1.37 |
| | Pneumonia, bacterial | **0.357** | 1.43 | - | | - | - | **0.375** | NS | 1.45 | 1.00 – 2.12 |
| | Rhythm disturbance | **-0.236** | 0.79 | - | | - | - | **-0.383** | NS | 0.68 | 0.41 – 1.14 |
| | Cardiovascular medical other | **-0.292** | 0.75 | - | | - | - | **-0.432** | NS | 0.65 | 0.36 – 1.17 |
| | COPD | **0.438** | 1.55 | - | | - | - | **0.108** | NS | 1.11 | 0.70 – 1.78 |
| | Sepsis | **0.354** | 1.42 | - | | - | - | **0.009** | NS | 1.01 | 0.65 – 1.57 |
| | Drug intoxication / overdose | **-1.528** | 0.22 | - | | - | - | **-1.286** | * | 0.28 | 0.09 – 0.83 |
| | Intracranial hemorrhage | **1.521** | 4.58 | - | | - | - | **1.426** | * | 4.16 | 2.49 – 6.97 |
| | Diabetic ketoacidosis | **-1.920** | 0.15 | - | | - | - | **-1.797** | * | 0.17 | 0.05 – 0.57 |
| | Other diagnoses‖ | **n/a** | n/a | - | | - | - | **0.771** | * | 2.16 | 1.63 – 2.86 |

† = Diagnostic category coefficients were NOT re-estimated
‡ = Diagnostic category coefficients were re-estimated
*** = Coefficient was significant (P value < 0.05) in logistic regression**

§ = Relative contribution of the diagnostic categories as a whole in predicting mortality compared to the other variables in the risk equation.
‖ = Diagnostic categories with insufficient patients to generate a coefficient were combined into a single category

Fourteen of the 15 re-estimated coefficients in the $MPM_0$ II model were significant in the re-estimated model. The only coefficient found not to be significant was for gastrointestinal bleeding. The majority of the re-estimated coefficients were also not statistically different from those in the original model. There was, however, a two-fold increase in the coefficients for HR greater than 150 and for a history of CPR prior to admission. Additionally, acute diagnosis of cardiac dysrhythmia was found to be protective in the re-estimated model. One possible explanation for these finding is that there may be a lower threshold for admitting patients with this diagnosis to the ICU, so as a whole their risk of mortality for this group of patients was lower.

In the re-estimated SAPS II model, the coefficient for the SAPS II score was significant, but not the log of the SAPS II score. The re-estimated coefficients were not statistically significantly different from those of the original model. The re-estimated APACHE II models produced coefficients for emergency surgery that were significant in the re-estimated model two but not model one. The coefficient for the APACHE II score was significant in the both re-estimated models. Coefficients for the diagnostic categories varied, with only nine being found to be significant in model 2. Many of the coefficients in the APACHE III re-estimated models were not significant. Lymphoma and metastatic cancer were the only significant variables for past medical history. Pre-ICU length of stay and excess hospital length of stay were also shown to be significant variables in the re-estimated models. Of the diagnostic categories, only three diagnoses were significant in the re-estimated model: drug intoxication/overdose, intracranial hemorrhage, and diabetic ketoacidosis.

***Summary***

With re-estimation, all the models demonstrated good calibration. $MPM_0$ II had a statistically lower discrimination than the other models. While calibration can be improved with re-estimation of coefficients, discrimination cannot be improved without changing the model. In selecting the optimal model, the predictive performance of the models should be compared to their data collection burden. $MPM_0$ II is the least burdensome as it only requires 15 variables. While APACHE II and APACHE III require significantly more data collection than $MPM_0$II or SAPS II, all of the data may not be needed. In the re-estimated models, many of the coefficients of the variables used in the risk equation were not significant. This may imply that a simplified version of the APACHE models could be devised that has less data collection with comparable performance by eliminating these variables. Particularly relevant is whether the reason for admission (diagnostic category) could be simplified or eliminated. Of the 50 APACHE II diagnostic coefficients, only nine were found to be significant. In APACHE III, only three of the diagnostic coefficients were significant. This may be partially due to an insufficient sample size; however, our sample was comparable in size to the APACHE II dataset (5,030 patients). The re-estimated SAPS II model has similar discrimination and calibration to the APACHE models, but requires less data collection. Most importantly, SAPS II does not require the coding of a specific reason for admission. In conclusion, with re-estimation, the models all had comparably good performance with the exception of $MPM_0$ II, which demonstrated lower discrimination. With similar performance, determining the optimal model may be based in part on the data collection burden. While the APACHE models are more burdensome than SAPS II, it may be possible to exclude several variables from the APACHE models that were found to be non-significant and reduce their data burden without impacting their performance. $MPM_0$ II requires the least data collection burden, but it comes at the cost of lowered discrimination.

## Results III: Comparative Performance of the Hospitals

The ability of the risk-adjusted mortality models to assess the comparative performance of the hospitals was evaluated using standardized mortality ratios (SMR). For each model, the standardized mortality ratio (observed deaths / predicted deaths) and the 95% confidence interval of the ratio were generated for each individual hospital. Only hospitals with greater than 100 ICU patients meeting the inclusion and exclusion criteria for all four models were included. Table 11.1 identifies high and low outliers by each risk-adjusted model, including the re-estimated models. High outliers (higher than expected mortality) were defined as hospitals whose 95% confidence interval for their SMR was greater than 1. Conversely, low outliers (lower than expected mortality) were defined as hospitals whose 95% confidence interval for their SMR was lower than 1.

### Table 11.1
### Statistically significant high and low hospital outliers

| | | ORIGINAL | | | | Re-estimated Models (1) | | | | Models (2) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Hospital | N | MPM | SAPS | APII | APIII | MPM | SAPS | APII | APIII | AP II | AP III |
| A | 120 | L | L | L | L | L | | | | L | |
| B | 176 | L | L | L | L | L | L | L | L | L | L |
| C | 355 | L | L | L | L | L | | | | | |
| D | 326 | L | L | L | L | L | L | | | | |
| E | 311 | L | L | L | L | | L | L | | L | |
| F | 328 | L | L | L | | | | | | | |
| G | 170 | | L | L | | | | | | | |
| H | 188 | | L | L | | | | | | | |
| I | 213 | | | L | | | H | | | | |
| J | 320 | | | | | | H | H | | H | |
| K | 369 | L | L | L | | | | | | | |
| L | 268 | | L | L | | | | | | | |
| M | 167 | | L | L | | | | | | | |
| N | 177 | | | | | | | | | | |
| O | 196 | | | | | | H | H | | | |
| P | 441 | H | L | L | | H | | | | | |
| Q | 148 | H | H | H | H | H | H | H | H | H | H |

H = High Outlier ; L = Low Outlier

Since the original models are poorly calibrated and over-predict mortality, depending on the model, 29 -76% of the hospitals were identified as low outliers. Once the models were re-estimated, fewer hospitals were identified as outliers. The APACHE III model was the most conservative in predicting outliers (2) compared to APACHE II (5), MPM (6), and SAPS (7). The re-estimated models with the diagnostic coefficients re-estimated did not alter the outliers. Hospital B was identified as a low outlier irrespective of which model was used and Hospital Q as a high outlier.

Note, at the conventional p<.05 level, it would be expected to identify one hospital in 40 as a high mortality outlier and one in 40 as a low mortality outlier just by chance, even if all were of average quality. Since we do not have (at this point) other ways to validate whether a hospital that is an outlier based just on its outcomes is there by chance or because its processes are unusually good or poor (i.e., with respect to what clinical experts consider to be best practice) "outliers" cannot be labeled as either "good" or "bad," they merely have more or fewer than the expected number of deaths.

By model, the hospitals were also ranked by their SMR (see Table 11.2). The lowest SMR received a rank of 1 and the highest SMR a rank of 17. In addition, the ranks were coded by color.

## Table 11.2
## Rank order of hospitals' SMRs
**(sorted by MPM ranking using re-estimated model)**

## Legend:

| | |
|---|---|
| **TOP QUARTILE (Lowest SMR)** | **2nd QUARTILE (2nd Lowest SMR)** |
| 3rd QUARTILE (2nd Highest SMR) | BOTTOM QUARTILE (Highest SMR) |

| Hospital | Original Models | | | | Re-estimated Models (1)* | | | | Model (2)† | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MPM | SAPS | AP II | APIII | MPM | SAPS | APII | APIII | APII | APIII |
| A | 2 | 2 | 3 | 2 | 1 | 2 | 3 | 2 | 2 | 2 |
| B | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| C | 3 | 5 | 4 | 6 | 3 | 6 | 4 | 7 | 4 | 6 |
| D | 5 | 6 | 7 | 4 | 4 | 5 | 6 | 3 | 5 | 3 |
| E | 4 | 3 | 2 | 3 | 5 | 4 | 2 | 5 | 3 | 5 |
| F | 7 | 8 | 9 | 9 | 6 | 8 | 10 | 10 | 10 | 10 |
| G | 8 | 7 | 8 | 5 | 7 | 7 | 9 | 4 | 8 | 4 |
| H | 6 | 4 | 5 | 7 | 8 | 3 | 5 | 8 | 6 | 8 |
| I | 9 | 15 | 13 | 14 | 9 | 16 | 14 | 13 | 14 | 13 |
| J | 13 | 14 | 16 | 10 | 10 | 14 | 16 | 6 | 16 | 7 |
| K | 11 | 12 | 12 | 12 | 11 | 11 | 12 | 12 | 11 | 12 |
| L | 10 | 9 | 6 | 8 | 12 | 10 | 7 | 11 | 7 | 11 |
| M | 12 | 11 | 11 | 13 | 13 | 12 | 11 | 14 | 12 | 14 |
| N | 14 | 13 | 15 | 15 | 14 | 13 | 13 | 15 | 15 | 15 |
| O | 15 | 16 | 14 | 16 | 15 | 15 | 15 | 16 | 13 | 16 |
| P | 16 | 10 | 10 | 11 | 16 | 9 | 8 | 9 | 9 | 9 |
| Q | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 |

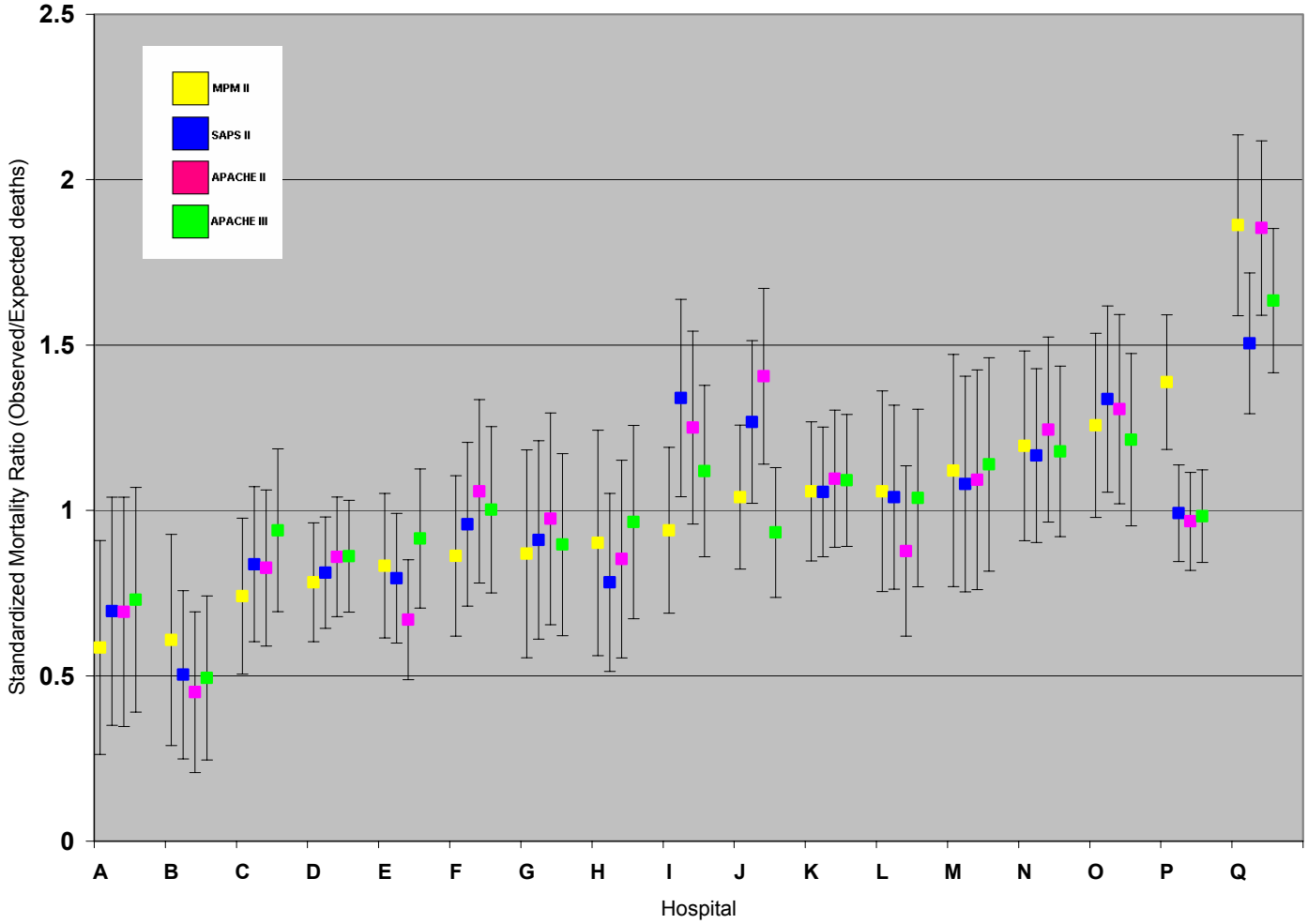
* = the re-estimated models of APACHE II and III, DO NOT include re-estimations of the diagnostic coefficients
† = the re-estimated models of APACHE II and III, DO include re-estimations of the diagnostic coefficients

In general, regardless of the model used, the best hospitals were ranked in the top quartile and the worst hospitals in the bottom quartile. When hospitals changed quartiles in a different model, generally they moved only a single quartile. The exceptions were Hospital I and P that

each moved two quartiles. No hospitals moved from the top quartile to the bottom quartile or from the bottom to the top.

A summary of the each hospital's SMR by model with 95% confidence interval is displayed in Table 11.3. The four models' point estimates of SMR for each hospital were not statistically different from each other with two exceptions. For Hospital J, the APACHE II and III SMR estimate were statistically different. In addition, the $MPM_0$ II estimate of SMR was statistically higher than the other three models for Hospital P. While the models may not differ statistically from each other, it is important to note the significant variability in the SMR (observed over predicted mortality) for the hospitals included in the study. The variability indicates that hospital performance reporting could be useful to improve ICU performance across California.

## Table 11.3
## SMR and 95% confidence interval of the hospitals
**(using all four of the re-estimated models)**



## CALICO 2004

While data from the first year report of the CALICO project suggests that ICU performance reporting may be practical and that existing risk-adjustment models must be recalibrated before they can be used, it is still not known what the most efficient approach to generating performance reporting will be. In its second year of data collection, CALICO plans to confirm the findings from the initial year of data collection as well as further explore the most accurate and efficient method to generate performance reports. The following is a list of the key tasks to be completed.

**I. Complete Recruitment and data collection**

CALICO has 27 hospitals recruited for the 2003-2004 data collection effort. Of these hospitals, 19 are trained and have begun to collect data. We cannot give the final number of hospitals for the 2003-2004 data collection effort until all the hospitals have submitted data. It has been our experience that some hospitals that commit to participation may drop out for various reasons anywhere in the process, even up to and during data collection.

The CALICO data collection tool was revised for the 2003-2004 data collection effort. More extensive data quality checks were added to the software along with some additional variables. The data collection tool was reformatted to improve the flow of information and ease data extraction. In spite of the additional variables, the reformatting should reduce the time needed for data input. More information about key variables was added to the face of the tool and input screen to increase accuracy and consistency of data. All manuals and codebooks were revised to assist in improving data quality. Training modules were updated and revised as needed.

**II. Assemble the final dataset for analyses**

The 2003-2004 dataset will contain approximately 6,000 additional patients, but may be larger depending on the ability of the hospitals to follow through on their commitments. We have made every effort to recruit hospitals that can and will complete data collection once in the project.

**III. Complete a 400 patient data quality audit**

We are completing a validation study including at least 10 patients from each hospital contributing patients to the dataset, for a minimum total of 400 patients. This is a retrospective study stratified by patient mortality status. We are sampling patients within each participating hospital, randomly selecting 7 survivors and 3 deaths from the cases submitted. Deaths are being over-sampled at twice the project death rate for ICU admissions during the study period (i.e., 30%). The over-sampling of deaths is intended to ensure that each hospital stratum contains enough high-risk patients who might be subject to abstraction errors or unmeasured risk factors and could, therefore, affect the predicted probability of death.

All ICU patient information used in this validation study is being re-abstracted from copies of the original patient medical record, either on site at the hospital or at the University of California, San Francisco. Participating hospitals are asked to provide a complete copy, or the original chart if on site, of each sampled patient's record. Registered nurses with experience in the ICU setting are trained to do a clinical abstraction, which averages 60 minutes a chart. The abstractors will be blinded to the results of the original chart abstraction done at the hospital. All abstractors have detailed data dictionaries, protocols, and ICU physician supervision. The study protocol has been approved by the Committee on Human Research at the University of California, San Francisco. All records are stored safely, not accessible to persons outside of the CALICO staff. The study is in compliance with both human subject and HIPAA regulations.

**IV. Complete the Comparison of Models, Evaluate New Models**

One of the key issues that needs to be explored in model comparison and development are the reasons for the poor calibration across most of the extant models. It is clear that case-mix differences can play an important role in this variation. Our strategy is to examine the models both before and after coefficient re-estimation using the CALICO dataset. We will then re-evaluate calibration, and finally add additional variables and interactions as necessary. We are currently beginning to explore the issues that need to be addressed before evaluating length of stay models.

In terms of model comparisons, our goals are to be able to understand how the models differ in terms of trade-offs between discrimination and calibration, to understand how they differ using a current California dataset when compared to the developers' datasets, and to develop a model that discriminates and calibrates well enough to be used for ICU hospital comparisons. Some of the issues that may affect the ability to develop models that work well are as follows:

a) It is possible that case-mix varies so markedly that ICUs need to be stratified, and models fitted for different types of ICUs; i.e. there is not "one" ICU case-mix nor a single model that can be applied to all ICUs.
b) The treatment that a patient receives in, for example, the emergency department before entering the ICU may affect some of the physiological variables that are collected at ICU admission. Thus, hospitals that stabilize patients in the emergency room before transfer to the ICU may cause the predicted mortality risk to be artificially lowered. The effect of treatment given in the emergency department, a common source of admission for ICU patients, may need to be considered.

**V. Seasonal Effects**

It has long been known that mortality rates and causes of mortality have an annual cycle, with rising mortality rates associated with respiratory illnesses occurring during each winter influenza season. However, it is not known whether this seasonal variation in mortality is significant in terms of reporting ICU performance. A priori, there are several reasons to think that it need not be. First, as long as all ICUs provide data covering roughly the same time period, it is unlikely that seasonal effects will have a differential impact on one ICU versus others, since all have the same exposure to the higher mortality period. Second, if one is using risk adjustment, it is

possible that, although the absolute mortality rate rises, this increase is corrected for by the risk adjustment model.

To reduce the impact of seasonality on the performance of ICU mortality models, each year's data collection was started on the same day, so that all hospitals were collecting data in approximately the same season. To further assess for evidence of significant seasonality, we will plot the O/E mortality rates for all CALICO participants by month of admission and month of discharge for the 2002 data. If there does not appear to be a seasonal effect, we will conclude the analysis there. If there is a seasonal effect, we will confirm this finding in the 2003-2004 data and correct for this trend in producing reports (by adjusting hospital performance in the appropriate direction based on month of the year).

## VI. ICD-9 Codes Model

We will obtain the discharge abstract ICD-9CM codes to determine how well a simple model using only variables in the discharge abstract could discriminate and calibrate relative to the customized extant models. We plan to use the OSHPD discharge abstract database from 2002 for this purpose. In addition, we intend to explore the option of using the data collected from the discharge abstract database in combination with the existing models or a subset of clinical variables contained within the models, to develop a new model with less data collection burden but comparable performance.