# California Intensive Care Outcomes (CALICO) Project
# Final Report

| | |
|---|---|
| Project Officer: | Denise Burgess |
| Project Organization: | Philip R Lee Institute for Health Policy Studies, University of California, San Francisco |
| Project Team: | R. Adams Dudley, MD, MBA, Principal Investigator<br>Mitzi Dean, MS, MHA<br>Eduard Vasilevskis, MD<br>Brian Cason, MD<br>Michael Kuzniewicz, MD<br>Ted Clay, MS<br>Rondall Lane, MD<br>Robert Chastain, PhD<br>Peter Bacchetti, PhD<br>Zita Konik, MD |

Grant No. RO1 HS13919-04<br>
Term of Grant: September 30, 2003 through March 23, 2009

This work was supported by the Agency for Healthcare Research and Quality.

## *Table of Contents*

## *Structured Abstract*

*Scope and Purpose*

Before the California Intensive Care Outcomes Project (CALICO), there was no direct comparison of updated ICU mortality models, or one common extant ICU length of stay (LOS) model (APACHE), and there was sparse reporting on mortality and good practice compliance correlations. The CALICO project's purpose is to provide new information across a suite of ICU measures to allow better evaluation of ICU measurement and reporting options.

*Methods*

Two retrospective patient record reviews were performed (11,300 initial and a 1,812 subset) of 2001-2004 discharges across 35 California hospitals. We calculated standardized mortality ratios (SMRs) per hospital using the MPM$_0$III, SAPS II, and ACACHE®IV models, comparing discrimination, calibration, data reliability, and abstraction time. We developed MPM$_0$III- and SAPS II-based models using mixed effects multi-level modeling and compared their accuracy with the APACHE®IV-LOS model. We compared the effect of the MPM$_0$II and APACHE III mortality models when examining good practice compliance, using logistic regression, across four conditions: myocardial infarction, community-acquired pneumonia, perioperative patients, and ventilator-dependent patients.

*Results*

The more recent APACHE and MPM models can be used for mortality, LOS (efficiency), and adjusted good practice measurement with similar results. The APACHE models had higher discrimination but three times the data burden.

*Key Words*: ICUs: outcome assessment (healthcare); quality of healthcare; risk adjustment; severity of illness index, length of stay assessment (healthcare)

## *Purpose of the Project*

The purpose of the project is to compare the performance of several models of Intensive Care Unit (ICU) mortality and length of stay and then examine whether complying with ICU good practices increases the survival rate of ICU patients after adjusting for initial mortality risk using various good practice measures and more than one risk adjustment model. The long-term goal of the principal investigator is to create performance reports that stimulate and provide the data for patient safety improvement. This project is focused on ICU performance and is intended to provide enough new information about ICU efficiency measurements, outcomes measurements, and good practice measure choices to assist policymakers and healthcare providers in choosing among viable alternative measurement options. The first part of the project, funded by the Office of Statewide Health Planning and Development (OSHPD), the Agency for Healthcare Research and Quality (AHRQ), and the Robert Wood Johnson Foundation (RWJF), compared the predictive accuracy and cost of data collection for the four commonly used ICU mortality risk adjustment models, a model using a state-wide administrative patient level database, enhancements of the five models, the development of an efficiency measure, ICU LOS, and comparison with an extant model. The second part of the project, funded by AHRQ and RWJF, compared the risk-adjusted survival rates of ICU patients who were given recommended good practices to those patients who did not receive them across four patient groups---myocardial infarction, community-acquired pneumonia, perioperative patients, and patients requiring mechanical ventilator assistance---using the two ICU mortality models that, as determined in the first part of the project, had the best mix of predictive accuracy and data collection burden. We also investigated whether good practice rates were correlated with risk-adjusted mortality rates across hospitals.

## *Scope of the Project*

### *Background*

The central hypothesis of this examination of the effects of ICU care is that information about patients' risk-adjusted mortality rates can be related to information about compliance with good practices of care across mortality models. The main analytic goals of the project are designed to produce better understanding of the strengths and weaknesses of commonly used ICU mortality and efficiency models and the relationship between the choice of mortality model and the choice of process measures. If it can be determined that, when using more than one robust ICU mortality model, there is significant variation across a diverse group of California hospitals and that poor performance in a currently accepted set of good practices is related to higher than expected mortality, then performance reports can be created with more assurance that they will provide a basis for improvements in patient care. In addition, more information will be available about the effect of the mortality model chosen and about the good practices being examined on good practice measurement. Comparing the performance of the extant risk-adjusted ICU length of stay model with a newly developed LOS model will provide needed information about whether the choice of a LOS model could potentially affect a hospital's comparative efficiency.

The work is timely because The Joint Commission (JC) is in the process of developing an ICU performance core measure set to be used for hospital accreditation, and the Hospital Quality Alliance has identified ICU care as the next arena (after surgical infection prophylaxis) for performance measurement implementation. The JC measure set will include a yet-to-be-determined ICU mortality risk adjustment model (CALICO results could influence that choice) and perhaps process measures currently being developed for and used in CALICO.

In California, the Office of Statewide Health Planning and Development is considering the public reporting of ICU process and mortality measures and has provided the public two reports based on the ICU mortality model development and comparisons that resulted from the early work of this project.[1]

### *Literature Review*

For more than 15 years, the healthcare community has known that the quality of healthcare delivered, as measured by outcomes and processes of care across hospitals, physician groups, and patients within a hospital, can vary dramatically[2] and that outcomes measurement alone, even when adequately applied, does not fully explain these differences.[3-7] Although ICU risk prediction models have been developed and updated,[8-13] before the CALICO project, there was no recent direct comparisons across the most commonly used models (SAPS, APACHE, and MPM) of their predictive accuracy in modern ICUs; their data collection burden had never been evaluated. Thus, current information needed to make an informed choice among ICU outcomes models, one of the important components of ICU performance measurement, was not available. Recent work on the application of mortality models indicates that, even with "good" risk adjustment, case-mix differences across hospitals may result in problems comparing some hospitals directly, depending on the breadth of the case-mix of the population when used for indirect standardization of mortality risk.[14] It is necessary, therefore, to use robust outcomes models and to use more than comparative outcomes when assessing most aspects of healthcare performance.

The minimum goal of most healthcare service providers is to ensure the provision of all necessary care (i.e., care that is generally considered to lead to benefits for the patient that outweigh the risks and that comprises the current standard of care for that condition while avoiding unnecessary, unsafe, inappropriate or inefficient care).[15, 16] Understanding whether a physician or institution has adequately met that goal is a complex, time-consuming quality measurement process. It usually requires, in part, the proper choice of outcome, risk adjustment for that outcome, measures of the services provided to patients and matching the measures with the appropriate outcomes. When the desire to compare one or more physicians or institutions is added, the task is further complicated by case-mix issues, possible structural differences, and the choice of appropriate benchmarks given those being compared. Fully understanding how patients' or hospitals' outcomes compare typically requires understanding the contribution of the Donabedian-defined structure, process, and outcomes categories for measuring healthcare quality.

This project was designed to increase the knowledge about how to best evaluate ICU performance. ICU performance is a worthy domain of study. Many types of patients are treated in the ICU, they are critically ill, and therefore the consequences of good or poor care are likely to have large impacts on patient outcomes.[17] From a policy perspective, ICU care is complex, employs some of the more advanced medical procedures and devices, has high nurse-to-patient ratios, and so is expensive to deliver.[18] Efficiency measurement can provide important assistance in understanding the balance between providing good care and containing costs. The portion of this study evaluating the relationship between process measures and outcomes is focused on ICU quality performance (i.e., determining if a patient received the type of care that is currently recommended for their specific set of healthcare issues and the relationship between that care and their survival rate). This examination of the effects of good practices employed will not determine if there are better care alternatives but rather will determine if those generally agreed upon are being implemented and are associated with mortality. As described by Donabedian, using the "integrative" outcome of risk-adjusted mortality necessitates looking at process, and often structural issues, to determine whether the patient outcomes are probabilistically related to the good practice compliance.[19]

Many factors can contribute to a patient's mortality in the hospital, and risk-adjusted mortality alone is usually not sufficient to understand how or if bench-marked mortality is related to the provision of quality care. Recently, three large studies (large number of ICUs, hospitals, and/or patients) found the effects of good practice implementation on risk-adjusted mortality ranged from a very significant effect to a significant but clinically small effect.[20-22] These three studies each used a single risk-adjustment model: 1) APACHE II, 2) a MEDPAR-based, condition-specific model, and 3) a model developed during the research project specific to patients with acute coronary syndrome. The large national Werner study, not limited to ICU patients, which examined the relationship between the 10 original process measures reported by those participating in the Hospital Compare public reports and hospital level risk-adjusted mortality, regrettably found that these performance measures only predicted small differences in hospital's risk-adjusted mortality rates.[20] The single condition ICU studies reported larger, clinically significant effects. It is difficult to determine if the different results from these studies are related to the risk adjustment model chosen, the measures chosen, the link between the two, the site (ICU or hospital), multi-condition versus one-condition project designs, real differences in the study populations, or other issues not measured.

Given the weak link shown in the Werner report between the quality metrics most universally being collected by hospitals at that time and the very common outcome metric, risk-adjusted mortality, the amount of public and non-public reporting being done using these variables is of concern. Reports comparing hospital performance are widespread and have a variety of uses. The Hospital Compare website (www.hospitalcompare.hhs.gov) currently provides a national inpatient hospital comparison for 29 measures of process information across five domains, three mortality outcomes (heart attack, heart failure, and pneumonia) and patient experiences with hospitals. Many states provide hospital comparative data based on data reported at both the federal and state levels.[23-26] In California, in addition to this information, there is a voluntary coalition of hospitals, health plans, consumers, state and federal organizations, and business representatives that are providing comparative acute hospital process and outcomes scores on more than 220 hospitals, including ICU processes and outcomes, to coalition members and the public (www.CalHospitalcompare.org). This dataset is also being used for "pay for performance" by at least one health plan in California. As consumers and providers compare hospital performance and health plans to make enrollment and payment decisions based on a broader range of data, it is important to have reliable efficiency models and understand whether or not the good practice and outcomes measures being reported can help identify hospitals that are more likely to adhere to good practices and to identify that the effect of these differences are clinically meaningful. The relationship between the risk models used, the good practice measures used, and the methods of comparing the relative contribution of each need to be better understood before they are used as a partial basis for payment decisions and more widely used for public reporting and for influencing insurers, providers of care, and consumers.

### *Goals of the Project*

The goals of the California Intensive Care Outcomes (CALICO) project were to assess the feasibility of, potential benefits from, and most efficient approach to, ICU performance reporting in California. The following is a description of the seven objectives of the project: 1) to evaluate the performance of $MPM_0$II & III, SAPS II, APACHE II, APACHE III & IV, and a model using the Patient Discharge Database (PDD) by applying them to a contemporary database (2001-2004) of California ICU patients, including an audit of the reliability of the model variables and customizing the models to the California dataset to improve their goodness-of-fit; 2) to use these models to determine whether there is significant variation among project hospitals in risk-adjusted mortality for ICU patients, and hence potential for measuring variability in quality of care;

3) 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; 4) to compare LOS models based on MPM, SAPS, and APACHE using the CALICO data; 5) to compare the effect of two common ICU mortality models on good practice measurement; 6) to investigate whether poor performance, judged by comparing the overall ICU risk-adjusted mortality rate between patients who received recommended practices and those who did not, indicates the presence of process of care patient safety deficiencies; and 7) to determine if there was a hospital-level correlation between condition-specific mortality and condition-specific good practices.

## *Methods*

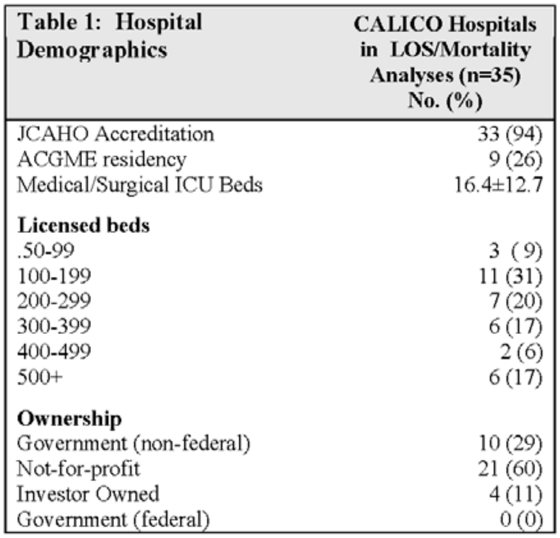
To achieve the project goals related to developing and evaluating updated, effective mortality models, we first used the 2006 American Hospital Association data to compare the CALICO hospital characteristics to all California hospitals with > 50 patients. We then evaluated the performance of $MPM_0$II & III, SAPS II, and APACHE II, III, & IV as specified by their developers and (in the case of APACHE® IV) by The Joint Commission,[8-10, 27-30] and we evaluated a model based on the PDD by applying them to a contemporary database (8/2001-9/2004) of 11,300 California ICU patients. We included a 400-patient audit of the reliability of the model variables and customized the models to the California dataset to improve their goodness-of-fit. Second, we used these models to determine whether there was significant variation among the final 35 project hospitals in risk-adjusted mortality for ICU patients, indicating potential for measuring variability in quality of care, and conducted supplemental analyses using the $MPM_0$III and APACHE® IV. Third, we compared the accuracy of three LOS models by calculating grouped coefficients of determination, assessing differences between observed and predicted LOS across subgroups and then assessed intra-class correlations of observed/expected LOS ratios between models using the CALICO patients. Fourth, we compared the available mortality models in terms of their predictive performance versus the burden of data collection and chose two models adequate for process measure evaluation. To achieve the goal of determining the relationship of the mortality model on good practice measurement, we used the $MPM_0$II and the APACHE III ICU mortality models to examine information abstracted from the charts of 1,802 patients (a subset of the CALICO patients) across four good practice conditions or treatment groups: 1) community-acquired pneumonia (CAP), 2) myocardial infarction (MI), 3) perioperative (periop), and 4) ventilator (vent). We included a 5% random sample re-abstraction to determine the reliability of the good practice variables. Fifth, to investigate whether condition-specific mortality performance was associated with the presence of condition-specific process of care deficiencies, we used logistic regression and condition-specific measures, one by one within conditions, controlling for ineligible patients and the variables in the APACHE III or $MPM_0$II. Next, we examined the effect of combining the measures within a condition (e.g., ventilator dependent patients); finally, we assessed the feasibility of looking at the effect of good practices on the hospital level.

### *The Comparison of ICU mortality and LOS models*

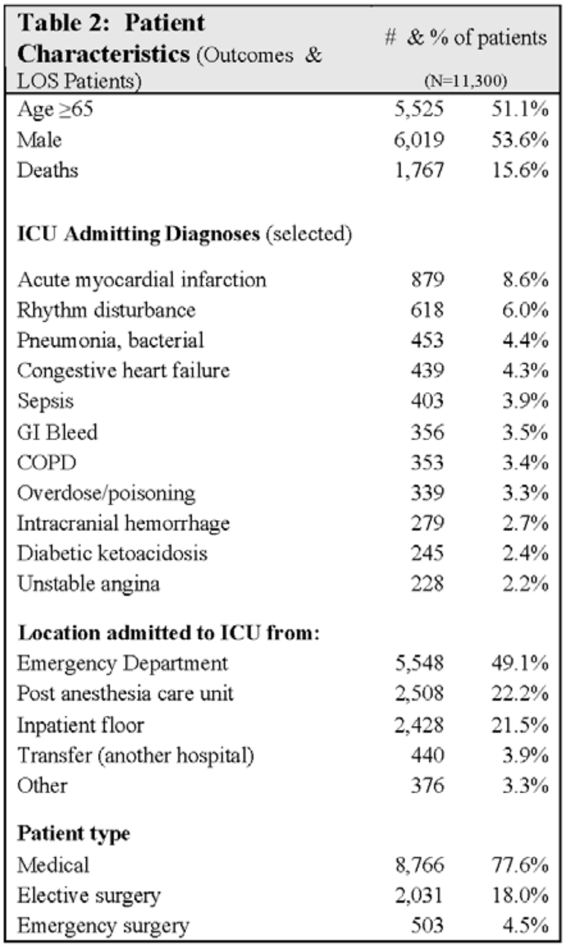
#### *Hospital Selection*

All California hospitals with a patient population of more than 50 patients were invited to join the portion of the study devoted to comparing mortality and LOS models.

Extensive recruitment activities included regional presentations on the study, calls to every hospital with an ICU, and presentations by the Principal Investigator at ICU-related conferences and meetings with the appropriate hospital staff, including the ICU physician in charge, ICU nurse managers, and quality improvement staff from the hospitals expressing interest. Thirty-five hospitals submitted data on 12,409 patients. We excluded 1,109 patients, as follows: 714 patients were readmissions to the ICU, 266 had excluded diagnoses, 62 had missing mortality model information, 43 had an ICU stay of < 4 hours, 24 were < 18 years old, and four were excluded from LOS analyses only. Tables 1 and 2 below describe hospital and patient characteristics of the 11,300 patients used to construct and compare the mortality and LOS of stay models.



| **Table 1: Hospital Demographics** | **CALICO Hospitals in LOS/Mortality Analyses (n=35) No. (%)** |
|---|---|
| JCAHO Accreditation | 33 (94) |
| ACGME residency | 9 (26) |
| Medical/Surgical ICU Beds | 16.4±12.7 |
| **Licensed beds** | |
| .50-99 | 3 ( 9) |
| 100-199 | 11 (31) |
| 200-299 | 7 (20) |
| 300-399 | 6 (17) |
| 400-499 | 2 (6) |
| 500+ | 6 (17) |
| **Ownership** | |
| Government (non-federal) | 10 (29) |
| Not-for-profit | 21 (60) |
| Investor Owned | 4 (11) |
| Government (federal) | 0 (0) |



| **Table 2: Patient Characteristics** (Outcomes & LOS Patients) | # & % of patients (N=11,300) | |
|---|---|---|
| Age ≥65 | 5,525 | 51.1% |
| Male | 6,019 | 53.6% |
| Deaths | 1,767 | 15.6% |
| **ICU Admitting Diagnoses** (selected) | | |
| Acute myocardial infarction | 879 | 8.6% |
| Rhythm disturbance | 618 | 6.0% |
| Pneumonia, bacterial | 453 | 4.4% |
| Congestive heart failure | 439 | 4.3% |
| Sepsis | 403 | 3.9% |
| GI Bleed | 356 | 3.5% |
| COPD | 353 | 3.4% |
| Overdose/poisoning | 339 | 3.3% |
| Intracranial hemorrhage | 279 | 2.7% |
| Diabetic ketoacidosis | 245 | 2.4% |
| Unstable angina | 228 | 2.2% |
| **Location admitted to ICU from:** | | |
| Emergency Department | 5,548 | 49.1% |
| Post anesthesia care unit | 2,508 | 22.2% |
| Inpatient floor | 2,428 | 21.5% |
| Transfer (another hospital) | 440 | 3.9% |
| Other | 376 | 3.3% |
| **Patient type** | | |
| Medical | 8,766 | 77.6% |
| Elective surgery | 2,031 | 18.0% |
| Emergency surgery | 503 | 4.5% |

*Patient Selection, Gender and Minority Inclusion, and Data Quality*

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. Although we did not specifically collect data on race or ethnicity, we accrued patients by enrolling all consecutive ICU patients over 18 years old from a wide range of hospitals by type. Members of vulnerable populations were included. Patients admitted to rule out myocardial infarction who were not found to have a critical illness were excluded. Data collection was proportional to reduce the burden on small hospitals and to explore the effect of severity and case-mix differences at larger hospitals, but statistical considerations required a minimum sample size of at least 200 patients per hospital for hospital-level analyses.

To assess the mortality models, demographic, clinical, and limited therapeutic data were collected by registered nurses through retrospective chart review of ICU patients from the 35 California hospitals that volunteered to join CALICO. Abstractors were instructed to collect all variables needed for the $\mathrm{MPM}_0$II & III, SAPS II, APACHE II, and APACHE III & IV on consecutive eligible patients and continue until their target sample size was reached.

Data quality was monitored throughout the project through initial and subsequent training of data collectors using in-person training, follow-up training materials and data dictionaries, automated data quality checks internal to the data collection software, and electronic screens applied to the data following data submissions. Physician support was available throughout the project. In addition, a 400-patient audit was conducted to allow calculation of inter-rater reliability statistics (percent agreement and kappas).

*Analyses*

For mortality model comparisons, mortality predictions were calculated for each patient using the six 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. We used logistic regression to re-estimate the coefficients.

When these ICU risk-adjustment models have been applied to populations distinct from the ones on which they were developed, each model has maintained adequate discrimination but shown poor calibration. To improve calibration, we used logistic regression to re-estimate the coefficients in the models using the CALICO ICU population. The methods used were similar to prior studies that customized the models to new populations[31-37] and are fully described in the report to the Office of Statewide Health Planning and Development on development of the ICU mortality models.[1] In addition, a simplified APACHE III model was developed using the APACHE III re-estimated model and then reclassifying each patient's reason for admission into one of the nine categories, eight by body system and one for overdose/poisoning (the APACHE III-System model).

Two models were developed that used variables available from the Patient Discharge Database (PDD), an administrative database reported to OSHPD. The first used as predictors only variables in the PDD: age, gender, primary reason for hospital admission, and other conditions present on hospital admission. The second model (PDD+) used these PDD data plus clinical variables that would be easy to collect via chart abstraction. Each of these clinical variables (heart rate, blood pressure, Glasgow coma score (GCS), need for mechanical ventilation, presence of an intracranial mass, and type of ICU admission) came from the $\mathrm{MPM}_0$ II model. To improve the calibration of the PDD+ model, heart rate, blood pressure and GCS were treated as continuous variables instead of being dichotomized as they were in $\mathrm{MPM}_0$ II. Over the course of the project, analyses were updated using the APACHE® IV and the $\mathrm{MPM}_0$III[8]; variables necessary to update to the SAPS III were not collected, as the updated model became available after data collection was completed. To compare the time needed to abstract data for these three models, three auditors used for the inter-rater reliability assessment abstracted data from 30 randomly selected patients. Each auditor alternated among the three models, so they were using the three models roughly equally across the 30 patients.

The performance of each hospital was evaluated using standardized mortality ratios (SMRs). The expected probability of mortality was calculated for each patient using the re-estimated coefficients from the ICU risk-adjustment models. To get an SMR for each hospital, the total observed mortality was divided by the model-specific expected mortality.

The ability of each of the models to identify hospital outliers was evaluated in several ways. The first was to determine whether the 95% confidence interval of the SMR included 1.0. The second approach involved a hospital fixed effect model. This method compares each hospital effect versus the un-weighted average of all the hospitals. A logistic regression is used to estimate the effect of each hospital on the overall model.  Finally, for each hospital, a "contrast" test between that hospital's effect and the average effect of all the hospitals was performed.

For LOS model development and comparison, data were again divided into development (60%) and validation samples, and Student's t-test, and Kruskal-Wallis tests and chi-squared tests were used to compare sample characteristics. LOS was calculated in days (to significant second digit) and truncated at 30 days. Mixed effects, multilevel modeling was used to generate ICU LOS prediction models for $MPM_0III$, APACHE® IV and SAPS II using re-estimated coefficients. Model performance was assessed in the validation sample using Student's t-tests to compare mean observed to mean predicted LOS for the entire population and for subgroups. Deciles of predicted LOS across the models were assessed using paired Student's t-tests and calibration curves. Model variance was determined using the $R^2$. To examine the proportion of variation across hospitals for hospitals with > 100 admissions, we used bivariate regressions of the mean observed LOS against the mean predicted LOS.[38] We standardized hospital's length of stay ratio (SLOSR) by dividing the mean observed hospital LOS by the mean predicted LOS. We then assessed intra-class correlations between SLOSRs produced by the three models to compare the model assessments of a hospital's LOS performance.

***Determining the relationship between mortality and process of care by condition and the effect of the choice of ICU model on this relationship***

To examine the relationship between condition-specific process of care patient safety deficiencies and condition-specific mortality performance indicators (i.e., risk-adjusted mortality) accounting for the known highly predictive risk models (APACHE III and $MPM_0$ II), we defined condition-specific patient safety deficiencies as noncompliance with good practice procedures in four Intensive Care Unit (ICU) patient conditions or treatment groups: 1) community-acquired pneumonia (CAP), 2) myocardial infarction (MI), 3) perioperative (periop), and ventilator dependent (vent). We chose the APACHE III and the $MPM_0II$ for comparison, because we found in our earlier work that the APACHE III provided the best discrimination of the extant models and calibration curves showed similar calibration across the deciles of risk, despite using variables in the model from the first 24 hours of ICU stay; the $MPM_0II$ uses only variables that are present on admission to the ICU and uses the fewest number of variables of the extant models.[7] Both models are widely used across the United States.[17, 39] The mortality performance outcome was death in the hospital (binary, whether death occurred while still in the hospital or not). For patient-level analyses, condition-specific good practice procedures at discharge were not evaluated in this study.

*Process Measure Selection*

Process measures were selected by a panel of two registered nurses with ICU experience and five physicians. The physician group included a pulmonary and critical care specialist, an anesthesiologist, and three physician fellows doing research on ICU performance: a hospitalist, a neonatologist, and an anesthesiologist. This panel chose the process measures after an extensive literature search and according to the recommendations in the following hierarchy: Joint Commission-recommended practice, professional societies' recommendations, large peer-reviewed studies or meta-analyses, and the physician panel.

Additional exclusions for the subset of process measures patients were as follows: MI patients were excluded if they were not admitted to the ICU from the emergency department or if the source of their MI was from atrial arrhythmia or was cocaine related. CAP patients were excluded if they were comfort care only, were transferred from another hospital, had been hospitalized within the past 14 days, or did not have one of the specified CAP diagnoses. For periop, patients were excluded if they were admitted for surgery following trauma or a coronary artery bypass graft or were participating in a clinical trial. Vent patients were excluded if they were not ventilated through a tracheostomy or endotracheal tube, did not have at least 24 consecutive hours of ventilation, and were admitted directly from the operating room or recovery room following surgery without discontinuation of mechanical ventilation.

*Hospital and Patient Selection*

After extensive presentations and related recruitment activities, 20 CALICO hospitals participated in the process measure portion of the study. For each hospital, we targeted 20 each MI and CAP and 40 each periop and vent patient abstractions, based on an analysis of the number of patients available by type in the original dataset. For hospitals that had more than the target number of patients in a type, a list of patients was randomly selected for process measure abstraction. All patients chosen for the process measure chart abstraction were a subset of the patients in the mortality portion of the study. Table 3 below describes the patient characteristics of the Process Measure analysis subset.

| **Table 3: Patient Characteristics (Process Measures)** | **# % All process pts** | **# % CAP pts** | **# % MI pts** | **# % Periop pts** | **# % Vent pts** |
|---|---|---|---|---|---|
| N Patients | 1812 | 263 | 234 | 696 | 619 |
| Age >= 65 | 52% | 56% | 66% | 47% | 50% |
| Male | 56% | 56% | 59% | 54% | 56% |
| Death Rate | 21% | 28% | 20% | 7% | 34% |
| Pre-ICU LOS - Mean days | 1.38 | 0.80 | 0.25 | 1.91 | 1.45 |
| Pre-ICU LOS - Median days | 0.33 | 0.26 | 0.21 | 0.47 | 0.24 |

*Data Collection and Data Quality*

Data collectors were registered nurses or physicians hired specifically for this project. Each data collector completed a training program that included a physician-led training on each tool using practice charts and a data dictionary. Successful completion of two practice chart re-abstractions in each of the four areas was followed by physician supervision throughout the chart abstraction process. Data abstraction was done by project staff onsite at each hospital, allowing use of electronic and hard copy information on each patient abstracted. Abstractions were reviewed and input by a registered nurse or a third-year medical student to allow another level of review. Charts were screened after input for inconsistencies or out of range values. These problems were resolved either through re-abstraction of the data or review of the hard copy forms. A 5% re-abstraction, including charts from each of the data collectors was completed to provide re-training and to determine the reliability of the process data. The process information on each patient was then linked to the mortality information available from the first portion of the study for further analyses. There were 11 patients dropped due to data inconsistencies or problems linking the data as follows: n=8 vent, n=1 CAP, and n=1 periop patient.

*Analyses*
All statistical tests were two sided. Statistical comparisons were performed at the 0.05 level of significance. No adjustments were made for multiple comparisons or to control for experiment-wise error rate. Patients missing Apache III model predicted probability of mortality had their $MPM_0II$ model predicted probability of mortality imputed as their Apache III and vice versa in most analyses. No other imputations were made for missing data.

*First*, within each indication (i.e., CAP, MI, periop, and vent), we examined each good practice one at a time, examining eligibility for the good practice measure and overall compliance among those who were eligible. We examined raw death rates among those who were eligible to receive the good practice as well as those who were ineligible.

*Second*, for each good practice, we ran statistical models adjusting for risk of death upon ICU admission.Whenever risk of death upon ICU admission was used, it was either the $MPM_0$ II or APACHE III probability of death based on patient characteristics at admission.*[1] In patients for whom either the MPM or the APACHE probability was missing, we imputed a value using the other non-missing value. In the adjusted analysis (Model 1), we used all patients in a logistic regression model predicting death using the following independent variables: 1) risk of death upon ICU admission, being either the $MPM_0$ II or APACHE III probability of death transformed as log(p/(1-p)), known as the logit transformation, 2) a dummy variable indicating that the patient was ineligible for the good practice, and 3) a dummy variable indicating that the patient was noncompliant. Thus, the reference value was the status of being both eligible and in compliance with the good practice. This model allowed the significance of the dummy variable for noncompliance to also be evaluated. For descriptive purposes, again working within an indication and a single good practice, we ran a separate logistic regression (Model 2), removing the dummy variable for noncompliance with good practice. These models produced the estimated probability of death, which was then summarized as the total predicted death rate in the good-practice group and the non-good-practice group. We tested for the significance of difference between the predicted mortality in the two groups using the Wilcoxon test. Within each group separately we calculated the observed/predicted mortality ratios. We calculated p-values for the effect of noncompliance from Model 1 to reflect the difference between these O/E ratios. For descriptive purposes, we also calculated the Raw minus Predicted death rates for each measure in each of the good practice groups.

*Third,* using a method published by Higashi, we created and analyzed an adjusted good practice score for each patient as a composite of a set of multiple robust good practice measures within a condition.[40] A good practice measure was considered robust if there were at least 50 patients in both the compliant and noncompliant groups. The score was calculated by dividing a patient's observed average compliance by his expected average compliance, looking only at the good practices for which the patient was eligible.

*[1] The MPM0II adjusts for three physiologic variables, three chronic diagnoses, six acute diagnoses, age, CPR prior to admission, mechanical ventilation, and nonelective surgery. The variables are collected at the time of admission or up to 1 hour after to the ICU. The APACHE III model attempts to capture the severity of the illness by illness type using the degree of abnormality of the patient's underlying physiology in the first 24 hours. At the time of this study, the APACHE III mortality prediction was determined by an equation including weights for a physiologic score, age, chronic health conditions, pre-ICU length of stay, location prior to admission, reason for ICU admission, and whether the patient had emergency surgery.

We calculated the expected average compliance for that patient as the mean of the overall compliance rates in all the good practices for which that patient was eligible. The effect of the adjusted good practice score (adjustment based on all measures within a condition for which the patient was eligible) was tested in a logistic regression model (Model 3) predicting death using the following independent variables: 1) risk of death upon ICU admission ($MPM_0$ II or Apache III) using the logit transformation and 2) the adjusted good practice score. The main result of interest was the statistical significance of the effect of the adjusted good practice score. An additional p-value was obtained by examining the distribution of the parameter estimates for the good practice score derived from 5,000 bootstrap samples of the patients with replacement. The reported two-tailed p-value was two times the mean rate at which the parameter estimates were found to be greater than or less than zero, whichever was smaller. This technique avoided some assumptions used in logistic regression in calculating the variance of parameter estimates.[41] Sensitivity and specificity of the models were examined using the C-statistic. A supplemental analysis was performed including "Aspirin in Hospital" in the adjusted score for MI patients.

*Fourth*, within the CALICO sample of ventilation patients, using hospitals with at least 10 patients, we examined the Pearson correlation coefficient for the relationship between the hospital mortality O/E ratios and the hospital mean of the (eligibility-adjusted) good practice scores. The O/E ratios were the ratio of observed to expected mortality rates, for which the expected mortality rate was the mean of the CALICO-derived probabilities of death (either MPM or APACHE).

## Results

### Determining the most cost efficient and effective ICU mortality models

The characteristics of CALICO hospitals did not differ significantly from those of all California hospitals in number of hospital beds by group size, percent hospitals with JC accreditation, ACGME residency, medical school affiliation, ownership, or number of medical/surgical ICU beds at the at the p-value < .05 level of significance. Inter-rater reliability was excellent across the physiological variables (agreement 91.5% to 98.8%, weighted k statistics ranging from 0.72 to 0.96). The Glasgow coma scale (86% agreement, k=0.55) showed good agreement. The APACHE reason for ICU admission was the least reliable variable (52.3% agreement; k=0.51). As shown in Table 4 below, the four re-estimated extant models showed adequate discrimination, with the APACHE III showing the highest discrimination (0.880, 95%CI 0.865-0.894) and the $MPM_0$II showing the lowest (0.811, 95% CI 0.791-0.830). The calibration was improved when compared to the models before re-estimation of the coefficients. The original poor fit across the four models (p-value<0.05) can be contrasted to the values in Table 4 below, although the HL statistics for the APACHE III still reached statistical significance. The large sample size of the validation dataset (> 3,000 patients) may have affected these results, as the HL statistic (p-value) is known to get smaller as a sample size increases. To further evaluate the calibration, calibration curves were produced. The APACHE III tended to overpredict death in a number of deciles, as did the $MPM_0$ II, but primarily in the highest deciles of risk. The SAPS II did not appear to have systematic bias in either direction, under- or over-predicting deaths. In later CALICO analyses using the APACHE® IV and the $MPM_0$ III, they demonstrated discrimination (0.892 (0.880-0.904) and 0.809 (0.791-0.826), respectively, similar to their earlier re-estimated versions. Although the APACHE® IV still had a higher HL statistic than the MPM or SAPS, calibration curves showed fit across the deciles of risk that was comparable to the other two models (not shown).

**Table 4: Discrimination and Calibration of Mortality Models**

| Model | AUC† (95% CI) | H-L‡ Statistic | |
|---|---|---|---|
| | | C Test | H Test |
| **MPM II** | | | |
| **Original** | 0.809 (0.789 – 0.828) | 52.9 (P<0.001) | 61.5 (P<0.001) |
| **Re-estimated Model** | 0.811 (0.791 – 0.830) | 11.3 (P=0.33) | 13.6 (P=0.19) |
| **SAPS II** | | | |
| **Original** | 0.870 (0.854 – 0.887) | 139.6 (P<0.001) | 143.5 (P<0.001) |
| **Re-estimated Model** | 0.870 (0.854 – 0.887) | 15.2 (P=0.12) | 6.9 (P=0.73) |
| **APACHE II** | | | |
| **Original** | 0.841 (0.823 – 0.859) | 155.0 (P<0.001) | 157.6 (P<0.001) |
| **Re-estimated Model** | 0.864 (0.848 – 0.879) | 15.2 (P=0.12) | 16.0 (P=0.10) |
| **APACHE III** | | | |
| **Original** | 0.881 (0.866 – 0.895) | 32.2 (P<0.001) | 37.3 (P<0.001) |
| **Re-estimated Model** | 0.880 (0.865 – 0.894) | 20.4 (P=0.026) | 27.1 (P=0.002) |

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

The final models, which examined the PDD and the PDD plus clinical data, had performance problems. The discrimination and calibration of the PDD models was inferior to the four models above (Discrimination – 0.774, 0.755 – 0.793, p<0.007 C test, p<0.001 H test). Because they were not currently in use and not superior to already existing models, we dropped them from additional consideration in the process measures comparisons.

The mean data abstraction times for the models tested were as follows: $MPM_0$ III, 11.1 min (95% CI, 8.7 to 13.4); SAPS II, 19.6 min (95% CI 17.0 to 22.2); and APACHE® IV, 37.3 min (95% CI 28.0 to 46.6 min). Differences were statistically significant at p <0.001.

Risk-adjusted mortality shows wide variation among project hospitals

The comparison of hospital risk-adjusted mortality across the models was quite consistent in all but one of the hospitals (i.e., the 95% confidence intervals for the point estimate of the SMR overlapped across the models). There was very significant variation among the hospitals, with a risk-adjusted mortality rate ranging from about 7% to 31% and an SMR of approximately 0.5 to 2.0 across five risk models ($MPM_0$ II, SAPS II, APACHE II and III, and the PDD used in this analysis). All models identified two high-mortality outliers and two low-mortality outliers using the SMR method. The results for fixed effects models and the contrast test were similar to the SMR results in terms of the number of outliers identified and the significance of hospital-specific effects. Supplemental model comparisons using the updated APACHE (IV) and MPM (III) models with the SAPS II showed very similar SMR results across the hospitals.[7]

In summary, there was enough evidence of variation in hospital mortality performance after risk adjustment across six extant ICU models to justify moving forward with examining the effects of good practice in the ICU on mortality after risk adjustment.

$MPM_0$ III-LOS and APACHE® IV-LOS more accurate than SAPS II-LOS for prediction of ICU LOS

Performance for each model was assessed in the 40% validation sample. Stratifying by age, APACHE® IV-LOS and $MPM_0$III each had a single age stratum with significant differences between observed and expected LOS. The SAPS II-LOS model underpredicted LOS for younger patients and overpredicted for older patients. Both APACHE® IV and $MPM_0$III accurately predicted ICU LOS for broad classifications of medical versus surgical reasons for admission to the ICU.

APACHE®IV-LOS was more accurate in more refined diagnostic categories. The APACHE®IV-LOS and the $MPM_0$III-LOS showed excellent fit after examining their calibration curves (not shown), but the SAPS II-LOS fit poorly across multiple deciles of risk, showing a significant difference ($p \leq 0.05$) between mean observed and predicted LOS in six deciles of predicted ICU LOS. The coefficients of determination for patient-level ICU LOS predictions were APACHE®IV-LOS ($R^2$=0.202), $MPM_0$III-LOS ($R^2$=0.098), and SAPS II-LOS ($R^2$=0.049). At the hospital level (grouped $R^2$), in an analysis of 29 CALICO hospitals with >100 admissions, $R^2$ s were as follows: APACHE®IV-LOS $R^2$=0.422, $MPM_0$III-LOS ($R^2$=0.279), and SAPS II-LOS ($R^2$=0.008). Again, as in the mortality work, there were substantial variations across hospitals in risk-adjusted LOS that did not seem to be related only to patients' severity of illness using the APACHE®IV-LOS and the $MPM_0$III-LOS models. The SAPS II was dropped from additional consideration due to poor calibration and low $R^2$.[42]

***Examining the relationship between condition-specific outcomes (mortality) and condition-specific processes of care***

*Data Quality*

In general, inter-rater reliability was good, with agreement ranging from 83% to 100% (12 of 17 were 100%) and weighted k statistics ranging from 0.67 to 1.00 across CAP, MI, and periop measures. Five of the seven ventilator measures also showed good agreement, ranging from 89% to 100%, with k statistics ranging from 0.75 to 0.89. The "waked" variable showed fair agreement at 74% agreement, with a kappa of 0.42. The most problematic variable was the composite variable SBT pure, which determined if a breathing trial was given after blood gases indicated that it was desirable to do so during a daily window between 10:00 am and 2:00 pm. We found that all components of this variable except time showed very good agreement (blood gas measurements indicate eligible for breathing trial (SBT), contraindications to SBT, whether or not a breathing trial was done, ineligible for SBT), ranging from 85% to 95% agreement, with k statistics ranging from 0.70 to 0.85. The composite measure as it was originally constructed, however, showed only 59% agreement and a k statistic of 0.19 if exact time matches for the start of the ventilator day were required. We are currently reconfiguring this variable to use alternative ventilator day definitions.

Condition-specific mortality, good practice eligibility, and compliance

The unadjusted mortality rate varied significantly across the four good process measure sets, as expected. Periop patients had the lowest raw mortality (7.5%), and ventilator-dependent patients had the highest (34.0%). In addition, the number of patients eligible for abstraction and subsequent analyses varied across the conditions, with periop and ventilator having the largest number of patients and, in the case of the ventilator patients, the most deaths. Every patient within a condition was eligible for at least one good practice, with the average number of measures per patient as follows: CAP 3.5 GPs, MI 6.9 GPs, periop 2.9 GPs, and vent 5.2 GPs. Good practice compliance ranged from a low of 14.1% on sputum culture before first antimicrobial in CAP patients to a high of 96.6% for beta blocker 24 hours after arrival in MI patients. The raw mortality rates for the ineligible patients (not shown) in any one good practice measure also varied across the measures, with a low of 6% for antibiotics before colon surgery and a high of 65% for waking sedated patients.

Comparison of predicted mortality and the effect of compliance (process performance) between good practice groups within condition by measure

When the difference between the raw mortality score and the predicted mortality score (both $MPM_0$ II and APACHE III) were examined across all measures and conditions, the raw mortality rate of the patients that got the good practices (GP) was 3.65%, on average, lower than their predicted rate.

Those who did not get the good practice (NGP) had a higher than predicted average mortality rate of 3.62%, a span of 7.27%. This overall trend varied by condition and by measure.

We next examined the predicted mortality of the GP and NGP groups (from Model 2) in an effort to understand any systematic differences in the risk of death at admission (and in the case of the APACHE III at admission and up to 24 hours after for some variables) to the ICU between these two groups. The average difference across all conditions and all measures, and using the information from both models, was 4.03% (GP 20.40% average predicted death rate (PDR), NGP 24.43% PDR). The test for the significance of these differences by measure within condition and by model shows a mixed result with no significant difference in seven of eight CAP measures (n=4 measures for each model), eight of 18 MI measures, four of 12 periop measures, and eight of 14 ventilator measures. We next reviewed the p-values of the logistic regression models (Model 1), which tested the significance of the effect of noncompliance by using all patients in a logistic regression model predicting death using the following independent variables: 1) risk of death upon ICU admission, being either the $MPM_0$ II or APACHE III probability of death transformed as log(p/(1-p)), known as the logit transformation; 2) a dummy variable indicating that the patient was ineligible for the good practice; and 3) a dummy variable indicating that the patient was noncompliant. The reference value was the status of being both eligible and in compliance with the good practice. Here, we found that the p-values were not significant for any of the periop or CAP measures, indicating no significant good practice effect on mortality risk, but varied by measure and to some extent by mortality model for the vent and MI patients (not shown).

*Effect of noncompliance with good practices on mortality risk using all robust good practice measures within a condition across two mortality models*

The Adjusted Good Practice score, which included a score for every patient in a condition across the measures for which they were eligible, was used in logistic regression models that also included the APACHE or the MPM risk of death (logit transformation). Consistent with the previous measure-by-measure analyses within CAP and periop, there was no significant good practice effect across the measures in these two conditions. Both the MPM and APACHE models showed a significant p-value for ventilator. For MI patients, the model was highly significant using the MPM and not significant using the APACHE. In a supplemental analysis that allowed the compliance with aspirin within 24 hours of hospital arrival into the Adjusted Good Practice score, both the MPM and the APACHE III had significant p-values.

*Exploration of hospital-level effects*

We examined only one condition, ventilator, for hospital-level good practice effects, as the ventilator measures had data from 17 of the 20 hospitals in the process measure subsample. We found, using the APACHE risk model, that there was a significant correlation between a hospital's mortality and good practice predictions, but the MPM correlation was not significant.

### *Discussion and Limitations*

Across three important components of ICU performance - ICU mortality, ICU length of stay, and the evaluation of the effects of good practice - both the APACHE and MPM models provide valuable information. Although the SAPS II model provides comparable information about mortality performance, it does not appear well suited when applied, as it is currently specified for mortality prediction, for use in LOS prediction.

We have shown some important differences that should be considered when using these current models to compare ICU performance. Across the three components examined in this study of quality measurement, ICU mortality, outcomes, and efficiency, the APACHE III and/or APACHE® IV appeared to have the best predictive ability. The APACHE III and IV outcomes and LOS predictive abilities are superior to the MPM II or III; on inspection, the calibration appears similar across the outcomes models, although the APACHE III and IV have significant p-values. If both LOS and outcomes are going to be examined, the SAPS II does not appear to provide useful LOS information, due to a very low coefficient of determination ($R^2$) at both the patient and hospital levels and poor calibration.[42]

At the hospital level, the prediction of mortality outliers is very similar using the SAPS II, APACHE III or IV, or MPM II or III, and LOS predictions using the APACHE IV or MPM III also produce similar outliers. The differences found in data collection burden, an average of 11.1 minutes for MPM and 37.3 minutes for APACHE® IV, and the fact that the $MPM_0II$ is the only model of the extant models that predicts from admission to the first hour only in the ICU, are significant considerations. The MPM and APACHE models also provided useful patient-level information within conditions about good practice compliance. If analyses are designed to look at hospital-level comparisons, then the lower cost MPM II or III data collection could be a viable choice.

*Limitations*: This study has several important limitations. First, in the mortality and LOS analyses, because we collected an unequal number of patients at each hospital to minimize the burden on smaller hospitals, these hospitals have a smaller chance of being labeled an outlier due to larger confidence intervals. We collected data over a 3-year period, and medical advances over this period may have resulted in decreased SMRs in the latter part of the data collection period. The sample of hospitals is volunteer and not random, so that could result in a nonrepresentative sample of hospitals, although our population had characteristics very similar to the overall population of California hospitals.[7] As it relates to the process portion of the study, this is a small sample of hospitals, and 20 hospitals did not allow a thorough examination of hospital-level effects. Furthermore, some potentially useful measures, although surely important, could not be evaluated due to small numbers. We compensated for this limitation by requiring at least 50 patients eligible for the measure in GP and NGP groups, but that necessarily eliminated some measures (fewer than 50) that appeared to be predictive when looked at singly. The retrospective nature of the study required the elimination of some potentially useful measures, such as head of bed elevation observations. The study design, retrospective chart abstraction, can identify associations only, not causality, and also introduces the issues of data completeness and potential confounders not measured.[44]

### *Summary*

We have shown a consistent pattern of predictive ability across the APACHE III and IV and the MPM II and III in both mortality and length of stay. The SAPS II model does not provide adequate prediction for a LOS model at the patient or hospital level. Both the APACHE III and the MPM II provided useful information about the effect of good practice compliance in some types of patients and again provided very similar results across the four conditions studied. The cost of collecting APACHE III or IV data should be considered when the analyses are going to be at the hospital level. The MPM II and APACHE III appear to predict similar results at the hospital level in both mortality and length of stay as well as patient-level results at the condition level.

## *Publications in print or expected from this grant*

Kuzniewicz, MW, Vasilevskis, EE, Lane, R, Dean, ML, Trivedi, NG, Rennie, DJ, Clay, T, Kotler, PK, Dudley, RA. Variation in ICU Risk-adjusted Mortality: Impact of Methods of Assessment and Potential Confounders. *CHEST*, 2008; 133(6):1319-27.

Vasilevskis, EE, Kuzniewicz, MW, Dean, ML, Clay, T, Rennie, DJ, Dudley, RA. Relationship between Discharge Practices and Intensive Care Unit In-hospital Mortality Performance: Evidence of a Discharge Bias. *Med Care*, in press.

Vasilevskis, EE, Kuzniewicz, MW, Dean, ML, Clay, T, Rennie, DJ, Vittinghoff, E, Dudley, RA. Mortality Probability Model III and Simplified Acute Physiology Score: Assessing their Value in Predicting Length of Stay and Comparison to APACHE IV. *CHEST*, in press.

Vasilevskis, EE, Kuzniewicz, MW, Dean, ML, Clay, T, Rennie, DJ, Dudley, RA. Predictors of Early Post Discharge Mortality in Critically Ill Patients: Lessons for Quality Performance and Quality Assessment. Submitted.

Dean, ML, Vasilevskis, EE, Cason, BA, Lane, R, Kuzniewicz, MW, Clay, T, Dudley, RA. Does ICU Mortality Performance predict ICU good practice performance? The effects of mortality models, conditions studied and the measures chosen. In preparation.

Cason, BA, Vasilevskis, EE, Dean, ML, Lane, R, Kuzniewicz, MW, Clay, T, Dudley, RA. Powerful predictors of ICU good practice: the importance of matching measures with outcomes. In preparation.

## *References*

1. California Office of Statewide Health Planning and Development. Healthcare outcomes page: California Intensive Care Outcomes (CALICO). (Accessed at http//www.oshpd.ca.gov/HID/Products/PatDischargeData/ICUDataCALICO.)

2. Stockwell DC, Slonim AD. Quality and safety in the intensive care unit. *J Intensive Care Med* 2006;21:199-210.

3. Brook RH, McGlynn EA, Cleary PD. Quality of health care. Part 2: measuring quality of care. *N Engl J Med* 1996;335:966-70.

4. Broder MS, Simon LP, Brook RH. Surgical quality: review of Californian measures. *BMJ* 2004;328:152-3.

5. Pronovost PJ, Jenckes MW, Dorman T, et al. Organizational characteristics of intensive care units related to outcomes of abdominal aortic surgery. *JAMA* 1999;281:1310-7.

6. Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Judging hospitals by severity-adjusted mortality rates: the influence of the severity-adjustment method. *Am J Public Health* 1996;86:1379-87.

7. Kuzniewicz MW, Vasilevskis EE, Lane R, et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. *CHEST* 2008;133:1319-27.

8. Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA. Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III). *Crit Care Med* 2007;35:827-35.

9. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985;13:818-29.

10. 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;100:1619-36.

11. 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;270:2478-86.
12. Metnitz PG, Moreno RP, Almeida E, et al. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. *Intensive Care Med* 2005;31:1336-44.
13. Moreno RP, Metnitz PG, Almeida E, et al. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005;31:1345-55.
14. Shahian DM, Normand SL. Comparison of "risk-adjusted" hospital outcomes. *Circulation* 2008;117:1955-63.
15. Schuster MA, McGlynn EA, Brook RH. How good is the quality of health care in the United States? *Milbank Q* 1998;76:517-63, 09.
16. Leape LL, Berwick DM. Five years after To Err Is Human: what have we learned? *JAMA* 2005;293:2384-90.
17. McMillan TR, Hyzy RC. Bringing quality improvement into the intensive care unit. *Crit Care Med* 2007;35:S59-65.
18. Berenholtz S, Pronovost P, Lipsett P, Dawson P, Dorman T. Assessing the effectiveness of critical pathways on reducing resource utilization in the surgical intensive care unit. *Intensive Care Med* 2001;27:1029-36.
19. Donabedian A. The role of outcomes in quality assessment and assurance. *QRB Qual Rev Bull* 1992;18:356-60.
20. Werner RM, Bradlow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. *JAMA* 2006;296:2694-702.
21. Ferrer R, Artigas A, Levy MM, et al. Improvement in process of care and outcome after a multicenter severe sepsis educational program in Spain. *JAMA* 2008;299:2294-303.
22. Peterson ED, Roe MT, Mulgund J, et al. Association between hospital process performance and outcomes among patients with acute coronary syndromes. *JAMA* 2006;295:1912-20.
23. Colorado Hospital Association. Colorado Hospital Report Card. (Accessed March 4, 2009, at http://www.cohospitalquality.org.)
24. Agency for Health Care Administration. FloridaHealthFinder.gov. (Accessed March 4, 2009, at http://www.floridahealthfinder.gov.)
25. State of New Jersey, Department of Health and Senior Services. Office of Health Care Quality Assessment. (Accessed March 4, 2009, at http://www.nj.gov/health/healthcarequality.)
26. California Office of Statewide Health Planning and Development (OSHPD). Healthcare Information Division. (Accessed March 4, 2009, at http://oshpd.ca.gov/HID/DataFlow/index.html.)
27. Le Gall JR LS, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993;270:2957-63.
28. Cerner Corporation. APACHE. (Accessed May 7, 2008, at www.cerner.com/public/MillenniumSolution.asp?id=3562.)
29. Joint Commission on Accreditation of Healthcare Organizations. National Hospital Quality Measures - ICU. (Accessed October 24, 2008, at http://www.jointcommission.org/PerformanceMeasurement/MeasureReserveLibrary/Spec+Manual+-+ICU.htm.)
30. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in intensive care units. *Ann Intern Med* 1993;118:753-61.

31. 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 II0. *Crit Care Med* 2002;30:1995-2002.

32. Markgraf R, Deutschinoff G, Pientka L, Scholten T, Lorenz C. Performance of the score systems Acute Physiology and Chronic Health Evaluation II and III at an interdisciplinary intensive care unit, after customization. *Crit Care* 2001;5:31-6.

33. 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* 1999;25:192-7.

34. Murphy-Filkins R, Teres D, Lemeshow S, Hosmer DW. Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: how to distinguish a general from a specialty intensive care unit. *Crit Care Med* 1996;24:1968-73.

35. 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;24:574-81.

36. Sirio CA, Shepardson LB, Rotondi AJ, et al. Community-wide assessment of intensive care outcomes using a physiologically based prognostic measure: implications for critical care delivery from Cleveland Health Quality Choice. *CHEST* 1999;115:793-801.

37. Zhu BP, Lemeshow S, Hosmer DW, Klar J, Avrunin J, Teres D. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: a simulation study. *Crit Care Med* 1996;24:57-63.

38. Zimmerman JE, Kramer AA, McNair DS, Malila FM, Shaffer VL. Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV. *Crit Care Med* 2006;34:2517-29.

39. Higgins TL, Teres D, Nathanson B. Outcome prediction in critical care: the Mortality Probability Models. *Curr Opin Crit Care* 2008;14:498-505.

40. Higashi T, Shekelle PG, Adams JL, et al. Quality of care is associated with survival in vulnerable older patients. *Ann Intern Med* 2005;143:274-81.

41. Efron BT, R.J. An Introduction to the Bootstrap: Chapman and Hall; 1993.

42. Vasilevskis E.E., Kuzniewicz M.W., Dean M.L., et al. Mortality Probability Model III and Simplified Acute Physiology Score: Assessing their value in predicting length of stay and comparison to APACHE IV. *CHEST* Revised and resubmitted.

43. Berenholtz SM, Dorman T, Ngo K, Pronovost PJ. Qualitative review of intensive care unit quality indicators. *J Crit Care* 2002;17:1-12.

44. Horn SD. Performance measures and clinical outcomes. JAMA 2006;296:2731-2.