NATIONAL QUALITY FORUM

Measure Evaluation 4.1
December 2009

This form contains the measure information submitted by stewards. Blank fields indicate no information was provided. Attachments also may have been submitted and are provided to reviewers. The subcriteria and most of the footnotes from the evaluation criteria are provided in Word comments within the form and will appear if your cursor is over the highlighted area. Hyperlinks to the evaluation criteria and ratings are provided in each section.

TAP/Workgroup (if utilized)**:** Complete all yellow highlighted areas of the form. Evaluate the extent to which each subcriterion is met. Based on your evaluation, summarize the strengths and weaknesses in each section.

Note: *If there is no TAP or workgroup, the SC also evaluates the subcriteria (yellow highlighted areas).*

Steering Committee: Complete all pink highlighted areas of the form. Review the workgroup/TAP assessment of the subcriteria, noting any areas of disagreement; then evaluate the extent to which each major criterion is met; and finally, indicate your recommendation for the endorsement. Provide the rationale for your ratings.

Evaluation ratings of the extent to which the criteria are met
C = Completely (unquestionably demonstrated to meet the criterion)
P = Partially (demonstrated to partially meet the criterion)
M = Minimally (addressed BUT demonstrated to only minimally meet the criterion)
N = Not at all (NOT addressed; OR incorrectly addressed; OR demonstrated to NOT meet the criterion)
NA = Not applicable (only an option for a few subcriteria as indicated)

(for NQF staff use) **NQF Review #:** 0703 **NQF Project:** Patient Outcomes Measures: Phases I and II
MEASURE DESCRIPTIVE INFORMATION
De.1 **Measure Title:** Intensive Care: In-hospital mortality rate
De.2 **Brief description of measure:** For all adult patients admitted to the intensive care unit (ICU), the percentage of patients whose hospital outcome is death; both observed and risk-adjusted mortality rates are reported with predicted rates based on the Intensive Care Outcomes Model - Mortality (ICOMmort).
1.1-2 **Type of Measure:** Outcome De.3 **If included in a composite or paired with another measure, please identify composite or paired measure**
De.4 **National Priority Partners Priority Area:** «npp_area» De.5 **IOM Quality Domain:** «quality_domain» De.6 **Consumer Care Need:** «consumer_care_need»

CONDITIONS FOR CONSIDERATION BY NQF	
Four conditions must be met before proposed measures may be considered and evaluated for suitability as voluntary consensus standards:	**NQF Staff**
A. The measure is in the public domain or an intellectual property (measure steward agreement) is signed. *Public domain only applies to governmental organizations. All non-government organizations must sign a measure steward agreement even if measures are made publicly and freely available.* A.1 **Do you attest that the measure steward holds intellectual property rights to the measure and the right to use aspects of the measure owned by another entity (e.g., risk model, code set)?** Yes A.2 **Indicate if Proprietary Measure** (*as defined in measure steward agreement*)**:** A.3 **Measure Steward Agreement:** Agreement will be signed and submitted prior to or at the time of measure submission A.4 **Measure Steward Agreement attached:**	**A** Y☐ N☐
B. The measure owner/steward verifies there is an identified responsible entity and process to maintain and	**B**

update the measure on a schedule that is commensurate with the rate of clinical innovation, but at least every 3 years. Yes, information provided in contact section	Y☐ N☐
C. The intended use of the measure includes both public reporting and quality improvement. ►**Purpose:** Public Reporting, Quality Improvement with Benchmarking (external benchmarking to multiple organizations)	**C** Y☐ N☐
D. The requested measure submission information is complete. Generally, measures should be fully developed and tested so that all the evaluation criteria have been addressed and information needed to evaluate the measure is provided. Measures that have not been tested are only potentially eligible for a time-limited endorsement and in that case, measure owners must verify that testing will be completed within 12 months of endorsement. D.1**Testing:** D.2 **Have NQF-endorsed measures been reviewed to identify if there are similar or related measures?** 0:Yes	**D** Y☐ N☐
(for NQF staff use) **Have all conditions for consideration been met?** **Staff Notes to Steward** (*if submission returned*)**:**	**Met** Y☐ N☐
Staff Notes to Reviewers (*issues or questions regarding any criteria*)**:**	
Staff Reviewer Name(s):	

TAP/Workgroup Reviewer Name:	
Steering Committee Reviewer Name:	
1. IMPORTANCE TO MEASURE AND REPORT	
Extent to which the specific measure focus is important to making significant gains in health care quality (safety, timeliness, effectiveness, efficiency, equity, patient-centeredness) and improving health outcomes for a specific high impact aspect of healthcare where there is variation in or overall poor performance. ***Measures must be judged to be important to measure and report*** *in order to be evaluated against the remaining criteria.* (evaluation criteria) **1a. High Impact**	**Eval Rating**
(for NQF staff use) **Specific NPP goal**:	
1a.1 **Demonstrated High Impact Aspect of Healthcare:** Patient/societal consequences of poor quality, Affects large numbers, Severity of illness, Frequently performed procedure, Leading cause of morbidity/mortality, High resource use 1a.2 1a.3 **Summary of Evidence of High Impact:** Between 1985-2000, the number of critical care beds in US acute care hospitals increased 26%. In addition, although the number of hospitals offering critical care has decreased, the proportion of total hospital beds assigned to critical care has increased. By 2005, critical care costs in the US were estimated to be $81.7 billion accounting for 13.4% of hospital costs, 4.1% of the national health expenditures and 0.66% of the gross domestic product. As a result, there is great interest in ensuring critical care is delivered at a high standard. One way in which the quality of ICU care can be measured is in-hospital mortality. As a quality indicator, mortality has been broadly accepted due to its relative ease of measurement and its importance to both clinicians and patients alike. Given the higher death rate of patients admitted to the ICU than those admitted to general hospital wards, mortality is an even more appropriate measure of outcome. In an effort to accurately predict risk of death, general ICU mortality risk-prediction models have been developed and refined over the last 20 years in order to better 'objectively' describe ICU patient populations and use these descriptive variables to estimate patients' risk for death. Patient characteristics, or case-mix, are significant contributors to risk of death, and risk-adjustment is prerequisite. 1a.4 **Citations for Evidence of High Impact:** Gunning K, Rowan, K. ABC of intensive care: outcome data	**1a** C☐ P☐ M☐ N☐

and scoring systems. BMJ 1999 Jul 24;317(7204):241-4.
Halpern NA. Can the costs of critical care be controlled? Curr Opin Crit Care 2009 Oct 9. [Epub ahead of print]
Halpern NA, Pastores SM et al. Changes in critical care beds and occupancy in the United states 1985-2000: differences attributable to hospital size. Crit Care Med 2006;34:2105-112.
Pronovost PJ, Miller MR et al. Developing and implementing measuers of quality of care in the intensive care unit. Curr Opin Crit Care 2001;7:297-303.
Rosenberg AL. Recent innovations in intensive care unit prediction models. Curr Opin Crit Care 2002;8:321-30.

1b. Opportunity for Improvement

1b.1 Benefits (improvements in quality) envisioned by use of this measure:

1b.2 Summary of data demonstrating performance gap (variation or overall poor performance) across providers:
A number of studies based on various large ICU patient samples have confirmed the marked variability in mortality outcomes. One of the earliest of such studies published in 1993 documented unadjusted in-hospital mortality rates from 42 intensive care units ranging from 6.4% to 40%, with 90% of this variation attributable to patient characteristics at admission. Corresponding observed to predicted mortality ratios varied from 0.67 to 1.25. Several studies have since observed similar wide variability among disparate samples. For instance, in a population of veterans, the mortality range was 2-30%, with observed to expected ratios ranging from 0.62-1.27. Again, in worldwide samples (as in the SAPS 3 database) and in geographically localized samples (California), mortality variability for patients admitted to the ICU persists.

1b.3 Citations for data on performance gap:
Knaus WA, Wagner DP, et al. Variations in mortality and length of stay in intensive care units. Ann Int Med 1993;118:753-61.
Kuzniewicz MW, Vasilevskis EE, Lane R et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 2008 Jun;133(6):1319-27.
Render ML, Kim M, Deddens J et al. Variation in outcomes in Veterans Affairs intensive care units with a computerized severity measure. Crit Care Med 2005;33(5): 930-9.
Rothen HU, Stricker K, Einfalt E et al. Variability in outcome and resource use in intensive care units. Intensive Care Med 2007;33:1329-36.

1b.4 Summary of Data on disparities by population group:
Among various populations, there are documented disparities in mortality outcomes following admission to the ICU. For instance, disease-specific racial variation has been noted among blacks, who have high mortality for critical care conditions such as sepsis or acute lung injury. In a similar fashion, the elderly have been looked at as a subpopulation in whom higher mortality rates occur after an ICU stay. In one study, older women fared worse than men in ICU outcomes. Age >65 together with mechanical ventilation >7d have together been demonstrated to be cofactors in predicting mortality. Aside from demographics, insurance status has also been implicated as a significant predictor, with the uninsured having higher risk of death.

1b.5 Citations for data on Disparities:
Barnato AE, Alexander SL et al. Racial variation in the incidence, care, and outcomes of severe sepsis: analysis of population, patient, and hospital characterisitics. Am J Respir Crit Care Med 2008 Feb 1;177(3):279-84.
Danis M, Linde-Zwirble WT et al. How does lack of insurance affect use of intensive care? A population-based study. Crit Care Med 2006 Aug;34(8):2235-6.
Erickson SE, Shlipak MH et al. Racial and ethnic disparities in mortality from acute lung injury. Crit Care Med 2009 Jan;37(1):1-6.
Feng Y, Amoateng-Adjepong Y et al. Age, duration of mechanical ventilation, and outcomes of patients who are critically ill. Chest 2009 Sept;136(3):157-64.
Fowler RA, Sabur N, Juurlink DN et al. Sex-and age-based differences in the delivery and outcomes of critical care. CMAJ 2007 Dec 4;177(12):1513-9.

1b
C☐
P☐
M☐
N☐

Yang Y, Yang KS et al. The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. J Crit Care 2009 Oct 14. [Epub ahead of print]

1c. Outcome or Evidence to Support Measure Focus

1c.1 Relationship to Outcomes (*For non-outcome measures, briefly describe the relationship to desired outcome. For outcomes, describe why it is relevant to the target population*)**:** Prevention of death is the reason why patients are admitted to the ICU. The critically ill are a uniquely vulnerable patient population, and given the high costs of their care, determining which hospitals have higher than expected rates of mortality following ICU admission is a meaningful measure of ICU performance.

1c.2-3. Type of Evidence: Randomized controlled trial, Observational study, Systematic synthesis of research

1c.4 Summary of Evidence (*as described in the criteria; for outcomes, summarize any evidence that healthcare services/care processes influence the outcome*)**:**
«outcomes_relationship_evidence_summary»

1c.5 Rating of strength/quality of evidence (*also provide narrative description of the rating and by whom*)**:**
Not-applicable

1c.6 Method for rating evidence: Not-applicable

1c.7 Summary of Controversy/Contradictory Evidence: In the process of evaluating possible structural or procedural factors contributing to mortality variation, a number of studies have been able to identify certain variables that are less significant. For instance, hospital size (>300 vs <300 patients) was not found to impact standardized mortality ratio in one particular study of 35 California hospital ICUs. Another study similarly looking at patient volume found that higher ICU patient volumes were associated with lower mortality rates, but only in high-risk critically ill adults. In a worldwide sample using the SAPS 3 database, authors found that factors related to nursing or physician staffing had no impact on performance ratings. Such findings may be specific to the populations to which the variables were studied, but warrant further examination.

1c.8 Citations for Evidence (*other than guidelines*)**:** Glance LG, Yue L et al. Impact of patient volume on the mortality rate of adult intensive care unit patients. Crit Care Med 2006;34(7):1925-34.
Kuzniewicz MW, Vasilevskis EE, Lane R et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest 2008 Jun;133(6):1319-27.
Rothen HU, Stricker K, Einfalt E et al. Variability in outcome and resource use in intensive care units. Intensive Care Med 2007;33:1329-36.

1c.9 Quote the Specific guideline recommendation (*including guideline number and/or page number*)**:**
Not-applicable

1c.10 Clinical Practice Guideline Citation: Not-applicable
1c.11 National Guideline Clearinghouse or other URL: Not-applicable

1c.12 Rating of strength of recommendation (*also provide narrative description of the rating and by whom*)**:**
Not-applicable

1c.13 Method for rating strength of recommendation (*If different from USPSTF system, also describe rating and how it relates to USPSTF)*:
Not-applicable

1c.14 Rationale for using this guideline over others:
Not-applicable

1c
C ☐
P ☐
M ☐
N ☐

TAP/Workgroup: What are the strengths and weaknesses in relation to the subcriteria for *Importance to Measure and Report?*	**1**
Steering Committee: Was the threshold criterion, *Importance to Measure and Report*, met? **Rationale:**	**1** **Y☐** **N☐**
2. SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES	
Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. (evaluation criteria)	**Eval Rating**
2a. MEASURE SPECIFICATIONS	
S.1 Do you have a web page where current detailed measure specifications can be obtained? **S.2 If yes, provide web page URL:** **2a. Precisely Specified**	
2a.1 Numerator Statement (*Brief, text description of the numerator - what is being measured about the target population, e.g. target condition, event, or outcome***):** Total number of eligible patients whose hospital outcome is death **2a.2 Numerator Time Window (***The time period in which cases are eligible for inclusion in the numerator***):** Not-applicable; Anyone with an ICU admission meeting eligibility criteria below is in the numerator. **2a.3 Numerator Details (***All information required to collect/calculate the numerator, including all codes, logic, and definitions***):** Eligible patients include those with an ICU stay of at least 4 hours and >18 years of age whose primary reason for admission does not include trauma, burns, or immediately post-coronary artery bypass graft surgery (CABG), as these patient groups are known to require unique risk-adjustment. Only index (initial) ICU admissions are recorded given that patient characteristics of readmissions are known to differ.	
2a.4 Denominator Statement (*Brief, text description of the denominator - target population being measured***):** Total number of eligible patients who are discharged (including deaths and transfers) **2a.5 Target population gender:** «population_gender» **2a.6 Target population age range:** «population_age» **2a.7 Denominator Time Window (***The time period in which cases are eligible for inclusion in the denominator***):** Not-applicable; Anyone with an ICU admission meeting eligibility criteria below is in the denominator. **2a.8 Denominator Details (***All information required to collect/calculate the denominator - the target population being measured - including all codes, logic, and definitions***):** Eligible patients include those with an ICU stay of at least 4 hours and >18 years of age whose primary reason for admission does not include trauma, burns, or immediately post-coronary artery bypass graft surgery (CABG), as these patient groups are known to require unique risk-adjustment. Only index (initial) ICU admissions are recorded given that patient characteristics of readmissions are known to differ.	
2a.9 Denominator Exclusions (*Brief text description of exclusions from the target population***):** <18 years of age at time of ICU admission, ICU readmission, <4 hours in ICU, primary admission due to trauma, burns, or immediately post-CABG, admitted to exclude myocardial infarction (MI) and subsequently found without MI or any other acute process requiring ICU care, transfers from another acute care hospital **2a.10 Denominator Exclusion Details (***All information required to collect exclusions to the denominator, including all codes, logic, and definitions***):** <18 years of age at time of ICU admission (with time of ICU admission abstracted preferably from ICU vital signs flowsheet), ICU readmission (i.e. not the patient´s first ICU admission during the current	**2a-specs** **C☐** **P☐** **M☐** **N☐**

hospitalization), <4 hours in ICU, primary admission due to trauma, burns, or immediately post-CABG, admitted to exclude myocardial infarction (MI) and subsequently found without MI or any other acute process requiring ICU care, patient transfers from another acute care hospital (i.e. patients whose physical site immediately prior to the index ICU admission was an acute care unit at an outside hospital)

2a.11 Stratification Details/Variables (*All information required to stratify the measure including the stratification variables, all codes, logic, and definitions***):**
Not-applicable

2a.12-13 Risk Adjustment Type: Stratification by risk category/subgroup

2a.14 Risk Adjustment Methodology/Variables (*List risk adjustment variables and describe conceptual models, statistical models, or other aspects of model or method***):**
Risk-adjustment variables include: age, heart rate >=150, SBP <=90, chronic renal, acute renal, GIB, cardiac arrhythmia, intracranial mass effect, mechanical ventilation, received CPR, cancer, cerebrovascular incident, cirrhosis, coma, medical admission or status post nonelective surgery, zero factor status (no risk factors other than age), and full code status (no restrictions on therapies or interventions at the time of ICU admission). The risk-adjustment model is based on the the Intensive Care Outcomes Model - Mortality (ICOMmort) with candidate interactions among variables and variable coefficients customized for the population of interest.

2a.15-17 Detailed risk model available Web page URL or attachment: Attachment MPMIII Model-633924315593551374.pdf

2a.18-19 Type of Score: Rate/proportion
2a.20 Interpretation of Score:
2a.21 Calculation Algorithm (*Describe the calculation of the measure as a flowchart or series of steps***):**
The hospital's observed mortality rate and risk-adjusted mortality rate are both calculated using the abstracted data. For each hospital, the model produces a median and 95% confidence interval for the Standardized Mortality Ratio (SMR), which is the death rate for the hospital adjusted to the average case mix.

2a.22 Describe the method for discriminating performance *(e.g., significance testing)***:**
«performance_discrimination_method»

2a.23 Sampling (Survey) Methodology *If measure is based on a sample (or survey), provide instructions for obtaining the sample, conducting the survey and guidance on minimum sample size (response rate)*:
the first 100 consecutive eligible patients per quarter

2a.24 Data Source (*Check the source(s) for which the measure is specified and tested)*
Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory, Paper Records

2a.25 Data source/data collection instrument (*Identify the specific data source/data collection instrument, e.g. name of database, clinical registry, collection instrument, etc.***):**
ICU Outcomes Data Collection Instrument

2a.26-28 Data source/data collection instrument reference web page URL or attachment: Attachment ICU Outcomes Tool-633924321207649804.pdf

2a.29-31 Data dictionary/code table web page URL or attachment: Attachment ICU Outcomes Data Dictionary-633924321323431795.pdf

2a.32-35 Level of Measurement/Analysis (*Check the level(s) for which the measure is specified and tested)*
Facility

2a.36-37 Care Settings (*Check the setting(s) for which the measure is specified and tested)*
Hospital/Acute Care Facility

2a.38-41 Clinical Services (*Healthcare services being measured, check all that apply*)
«clinical_services» «clinical_services_clinician_other» «clinical_services_behavorial_health_othe»

«clinical_services_other»

TESTING/ANALYSIS

2b. Reliability testing

2b.1 Data/sample *(description of data/sample and size)*: MPM III was developed on the Project IMPACT database of 124,855 patients treated in 135 ICUs at 98 hospitals between 10/2001 and 3/2004. The majority of the hospitals were in the United States. Although reliability of data collection was ensured by the developers of the MPM III model, analyses and statistics were not reported in their original publication (Crit Care Med 2007;35:827-35).

The subsequent reliability testing methods and results are taken from Chest 2008 Jun;133(6):1319-27, in which the data sample was comprised of 11,300 patients admitted to 35 California hospital ICUs from 2001-2004.

2b.2 Analytic Method *(type of reliability & rationale, method for testing)*:
Data were reabstracted by auditors on a 5% random sample of patients. Kappa statistics were calculated for interrator variability between the data abstractor and the auditor. The auditors were clinical nurses who were trained by the authors and completed extensive sample chart abstraction.

2b.3 Testing Results *(reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
For physiologic variables of the MPM III mortality prediction model, interrator reliability was excellent, with agreement ranging from 91.5 to 98.8%, and weighted kappa statistics ranging from 0.72-0.96.

2b C☐ P☐ M☐ N☐

2c. Validity testing

2c.1 Data/sample *(description of data/sample and size)*: 50,307 (40.3%) of the patients were used for model validation.

2c.2 Analytic Method *(type of validity & rationale, method for testing)*:
Model performance in the validation sample was tested using area under the receiver operating characteristics curve (ROC) >=0.75 as a measure of discrimination and the Hosmer-Lemeshow goodness-of-fit statistic as a measure of calibration. Acceptable calibration was defined as: a) a nonsignificant Hosmer-Lemeshow value; b) a Hosmer-Lemeshow decile calibration plot with a slope and intercept not differing significantly from 1 and 0, respectively; and c) the SMR on the validation set between 0.95-1.05 with its confidence intervals including 1.

2c.3 Testing Results *(statistical results, assessment of adequacy in the context of norms for the test conducted)*:
Area under the ROC curve was 0.823 (95% CI 0.818-0.828), the Hosmer-Lemeshow statistic 11.62 (p = 0.31), and the standardized mortality ratio was 1.018 (95% CI 0.996-1.040). Actual mortalities closely tracked MPM III predictions by deciles of predicted risk.

2c C☐ P☐ M☐ N☐

2d. Exclusions Justified

2d.1 Summary of Evidence supporting exclusion(s):
«exclusions_evidence»

2d.2 Citations for Evidence:
«exclusions_citations»

2d.3 Data/sample *(description of data/sample and size)*: Not-applicable

2d.4 Analytic Method *(type analysis & rationale)*:
Not-applicable

2d.5 Testing Results *(e.g., frequency, variability, sensitivity analyses)*:

2d C☐ P☐ M☐ N☐ NA☐

Not-applicable	
2e. Risk Adjustment for Outcomes/ Resource Use Measures **2e.1 Data/sample** *(description of data/sample and size)*: After 200 records were excluded due to missing essential MPM or outcome variables, 124,885 patients were available for analysis and model development. **2e.2 Analytic Method** *(type of risk adjustment, analysis, & rationale)*: Data analyzed were from ICUs with >= 100 patient records in the Project IMPACT database. The MPM II variables were used in the MPM III update, and new candidate variables were also evaluated. Univariate analysis assessed the relationship of the MPM II independent variables to mortality using Student's t-tests and chi-squared tests with a significance level of 0.05. Multivariate logistic regression with robust variance estimators was performed using variables with a significant univariate relationship to outcome. Interactions were considered (particularly between age and other MPM variables) since initial calibration suggested that age effects were influenced by the presence of comorbidities. **2e.3 Testing Results** *(risk model performance metrics)*: Not-applicable **2e.4 If outcome or resource use measure is not risk adjusted, provide rationale:** Not-applicable	**2e** C☐ P☐ M☐ N☐ NA☐
2f. Identification of Meaningful Differences in Performance **2f.1 Data/sample from Testing or Current Use** *(description of data/sample and size)*: contemporary California hospital sample of nearly 70,000 patients from 188 ICUs **2f.2 Methods to identify statistically significant and practically/meaningfully differences in performance** *(type of analysis & rationale)*: In order to evaluate the performance of the customized MPM III model in our population of interest, the AUROC was calculated and found to be 0.829 (a value comparable to that found by model developers.) Moreover, when hospital SMR rankings using MPM III were compared to those generated using MPM II, the Spearman rank correlation coefficient was 0.97, indicating a strong correlation in hospital rankings (i.e. performance measurement). **2f.3 Provide Measure Scores from Testing or Current Use** *(description of scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*: Using the customized version of the MPM III on the current database of California hospitals yielded the following summary statistics of the hospital SMRs: Mean 0.989293 Median 0.988061 Std Deviation 0.26515 Interquartile Range 0.83-1.13	**2f** C☐ P☐ M☐ N☐
2g. Comparability of Multiple Data Sources/Methods **2g.1 Data/sample** *(description of data/sample and size)*: Not-applicable **2g.2 Analytic Method** *(type of analysis & rationale)*: Not-applicable **2g.3 Testing Results** *(e.g., correlation statistics, comparison of rankings)*: Not-applicable	**2g** C☐ P☐ M☐ N☐ NA☐
2h. Disparities in Care **2h.1 If measure is stratified, provide stratified results** *(scores by stratified categories/cohorts)*: This measure is not stratified. **2h.2 If disparities have been reported/identified, but measure is not specified to detect disparities,**	**2h** C☐ P☐ M☐ N☐ NA☐

provide follow-up plans: Race/ethnicity could be added as a variable in the data collection tool (though it is not in the current tool). Results could easily be stratified if this variable was added.	
TAP/Workgroup: What are the strengths and weaknesses in relation to the subcriteria for *Scientific Acceptability of Measure Properties?*	**2**
Steering Committee: Overall, to what extent was the criterion, *Scientific Acceptability of Measure Properties*, met? **Rationale:**	***2*** **C☐** **P☐** **M☐** **N☐**
3. USABILITY	
Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making. (evaluation criteria)	**Eval Rating**
3a. Meaningful, Understandable, and Useful Information **3a.1 Current Use:** **3a.2 Use in a public reporting initiative (disclosure of performance results to the public at large)** (*If used in a public reporting initiative, provide name of initiative(s), locations, Web page URL(s). If not publicly reported, state the plans to achieve public reporting within 3 years*)**:** UCSF supports the use of this measure by the California Hospital Assessment and Reporting Taskforce (CHART). It is publicly reported at www.calhospitalcompare.org. **3a.3 If used in other programs/initiatives (***If used in quality improvement or other programs/initiatives, name of initiative(s), locations, Web page URL(s). If not used for QI, state the plans to achieve use for QI within 3 years***):** The MPM III model is used in Project IMPACT, a program of the Society for Critical Care Medicine, but that data is not publicly accessible. **Testing of Interpretability** (*Testing that demonstrates the results are understood by the potential users for public reporting and quality improvement*) **3a.4 Data/sample** *(description of data/sample and size)***:** «interpretability_data» **3a.5 Methods** *(e.g., focus group, survey, QI project)***:** «interpretability_method» **3a.6 Results** *(qualitative and/or quantitative results and conclusions)***:** «interpretability_results»	**3a** **C☐** **P☐** **M☐** **N☐**
3b/3c. Relation to other NQF-endorsed measures **3b.1 NQF # and Title of similar or related measures:**	
(for NQF staff use) **Notes on similar/related endorsed or submitted measures:**	
3b. Harmonization If this measure is related to measure(s) already endorsed by NQF (e.g., same topic, but different target population/setting/data source or different topic but same target population): **3b.2 Are the measure specifications harmonized? If not, why?**	**3b** **C☐** **P☐** **M☐** **N☐** **NA☐**
3c. Distinctive or Additive Value **3c.1 Describe the distinctive, improved, or additive value this measure provides to existing NQF-endorsed measures:** «similar_endorsed_measures_added_value» **5.1 If this measure is similar to measure(s) already endorsed by NQF (i.e., on the same topic and the**	**3c** **C☐** **P☐** **M☐** **N☐** **NA☐**

same target population), Describe why it is a more valid or efficient way to measure quality:	
TAP/Workgroup: What are the strengths and weaknesses in relation to the subcriteria for *Usability?*	**3**
Steering Committee: Overall, to what extent was the criterion, *Usability*, met? **Rationale:**	**3** **C☐** **P☐** **M☐** **N☐**
4. FEASIBILITY	
Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement. (evaluation criteria)	**Eval Rating**
4a. Data Generated as a Byproduct of Care Processes **4a.1-2 How are the data elements that are needed to compute measure scores generated?** Data generated as byproduct of care processes during care delivery (Data are generated and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition), Coding/abstraction performed by someone other than person obtaining original information (E.g., DRG, ICD-9 codes on claims, chart abstraction for quality measure or registry)	**4a** **C☐** **P☐** **M☐** **N☐**
4b. Electronic Sources **4b.1 Are all the data elements available electronically?** (*elements that are needed to compute measure scores are in defined, computer-readable fields, e.g., electronic health record, electronic claims*) No **4b.2 If not, specify the near-term path to achieve electronic capture by most providers.** There already exist electronic medical record options that hospitals could purchase that would collect this data. However, most hospitals have not yet purchased such software.	**4b** **C☐** **P☐** **M☐** **N☐**
4c. Exclusions **4c.1 Do the specified exclusions require additional data sources beyond what is required for the numerator and denominator specifications?** «exclusions_additional_data_sources» **4c.2 If yes, provide justification.** «exclusions_additional_data_sources_justi»	**4c** **C☐** **P☐** **M☐** **N☐** **NA☐**
4d. Susceptibility to Inaccuracies, Errors, or Unintended Consequences **4d.1 Identify susceptibility to inaccuracies, errors, or unintended consequences of the measure and describe how these potential problems could be audited. If audited, provide results.** The potential unintended consequence is that hospitals may seek to avoid high-risk patients. One could monitor this behavior by evaluating changes in hospitals' risk-profiles over time.	**4d** **C☐** **P☐** **M☐** **N☐**
4e. Data Collection Strategy/Implementation **4e.1 Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data/missing data, timing/frequency of data collection, patient confidentiality, time/cost of data collection, other feasibility/ implementation issues:** In 188 hospitals in California (from small rural hospitals to the largest teaching hospitals), we have successfully collected this data. The average time per chart for an experienced data collector is 11-15 minutes. We collect data on 100 patients per quarter to minimize the data collection burden while still getting sufficient sample size to get precise estimates of hospital performance. However, an alternative target sample size could easily be chosen by users.	**4e** **C☐** **P☐** **M☐** **N☐**

4e.2 **Costs to implement the measure** (*costs of data collection, fees associated with proprietary measures*)**:** «implementation_costs» 4e.3 **Evidence for costs:** «implementation_costs_evidence» 4e.4 **Business case documentation:** «business_case»	
TAP/Workgroup: What are the strengths and weaknesses in relation to the subcriteria for *Feasibility?*	**4**
Steering Committee: Overall, to what extent was the criterion, *Feasibility*, met? **Rationale:**	**4** **C☐** **P☐** **M☐** **N☐**
RECOMMENDATION	
(for NQF staff use) **Check if measure is untested and only eligible for time-limited endorsement.**	**Time-limited ☐**
Steering Committee: Do you recommend for endorsement? **Comments:**	**Y☐** **N☐** **A☐**
CONTACT INFORMATION	
Co.1 **Measure Steward (Intellectual Property Owner)** Co.1 **Organization** «steward_intellectual_property_organizati» Co.2 **Point of Contact** R. Adams, Dudley , MD, MBA , adams.dudley@ucsf.edu , 415-476-8617-	
Measure Developer If different from Measure Steward Co.3 **Organization** Philip R. Lee Institute for Health Policy Studies, 3333 California Street, Suite 265, San Francisco , 94118 Co.4 **Point of Contact** R. Adams, Dudley , MD, MBA , adams.dudley@ucsf.edu , 415-476-8617-	
Co.5 **Submitter If different from Measure Steward POC** R. Adams, Dudley , MD, MBA , adams.dudley@ucsf.edu , 415-476-8617-, Philip R. Lee Institute for Health Policy Studies	
Co.6 **Additional organizations that sponsored/participated in measure development**	
ADDITIONAL INFORMATION	
Workgroup/Expert Panel involved in measure development Ad.1 **Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.** Not-applicable	
Ad.2 **If adapted, provide name of original measure:** Not-applicable Ad.3-5 **If adapted, provide original specifications URL or attachment** «adapted_original_measure_available» «adapted_original_measure_url_password» «adapted_original_measure_provided»	
Measure Developer/Steward Updates and Ongoing Maintenance Ad.6 **Year the measure was first released:** 1990 Ad.7 **Month and Year of most recent revision:** 03, 2007	

Ad.8 What is your frequency for review/update of this measure? annually **Ad.9 When is the next scheduled review/update for this measure?** 01, 2010
Ad.10 Copyright statement/disclaimers: is in the public doman
Ad.11 -13 Additional Information web page URL or attachment: «additional_information_available» «additional_information_url_password» «additional_information_provided»
Date of Submission (*MM/DD/YY*)**:** 06/13/2011

SUMMARY OF NQF-ENDORSED INTENSIVE CARE OUTCOMES MODELS FOR RISK ADJUSTED MORTALITY AND LENGTH OF STAY ($ICOM_{mort}$ and $ICOM_{LOS}$)

BACKGROUND

Importance of the ICU

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 rate reported ranging from 8-19%, or about 500,000 deaths annually.[1-3] The ICU is also one of the sites in which medical errors are most likely to occur because of the complexity of care.[4,5] 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.[6,7] In addition to its impact on mortality, critical care is a costly component of the national health care budget, with costs estimated to be $81.7 billion by 2005, accounting for 13.7% of hospital costs, 4.1% of national health expenditures, and 0.66% of the gross domestic product.[8] These costs are largely explained by the length of stay (LOS) in the ICU.[9,10] For these reasons, there has been substantial interest in measuring ICU outcomes, both in terms of mortality and resource utilization.

Variation in ICU Mortality and Resource Utilization

Considerable variation in mortality has also been observed among ICU patients, which persists even after adjustment for patient characteristics present at admission.[11-16] Twofold to threefold differences in ICU risk adjusted mortality that were previously reported[12] are still present in modern ICUs, irrespective of the model that is used to adjust for patient severity of illness.[13] Similar variation has also been seen in ICU length of stay (LOS), again even after accounting for patient risk factors.[16-19]

Extant ICU Outcome Risk Adjustment Models up to 2011

Clinicians and researchers have long recognized how important ICU performance is to overall hospital mortality. A significant amount of work has already been done to develop tools to assess ICU performance. The three most widely used general mortality risk adjustment models are the Mortality Probability Model (MPM), the Acute Physiology and Chronic Health Evaluation, and the Simplified Acute Physiology Score (SAPS). In a comparative analysis of these models, it has been shown that a hospital's mortality performance assessed by a standardized mortality ratio is not much impacted by which model is chosen.[17]

Of note, all models have been shown to need frequent recalibration.[20-24] That is, while it is reasonable to continue using the same clinical variables, the coefficients on those variables, and the possibility of interactions among the variables needs to be evaluated frequently (and whenever the models are applied to a new population).[25-30]

In addition, it has been shown that the variables used in each of these models can be incorporated into risk adjustment systems for ICU length of stay (LOS).[17] Again, the choice of model among the MPM, SAPS, and APACHE systems did not have much impact on each hospital's performance assessment.

There is, however, an important and stable difference between the models, and this relates to the data collection burden. Among randomly selected patients, the average time for chart abstraction for the MPM, SAPS, and APACHE models was 11, 20, and 37 minutes, respectively.[13]

In summary, the prior literature suggests that choice of model has little impact on hospital performance assessments, but major impact on data collection costs, with MPM being by far the least burdensome. For these reasons, we have recommended that models incorporating the MPM variables be used as the primary method of risk adjustment of ICU outcomes and we submitted such models to the National Quality Forum for evaluation.

Summary

Based on the clinical and economic significance of the ICU and the evidence that ICU performance varies, the National Quality Forum has endorsed measures of ICU outcomes (risk adjusted mortality and length of stay) for public reporting. In this document, we describe the requests made by the critical care community during evaluation of these models by the National Quality Forum and subsequently by the Hospital Quality Alliance. We then explain how we have adjusted our models—which we now refer to as the Intensive Care Outcomes Models ($ICOM_{mort}$ and $ICOM_{LOS}$ for ICU mortality and ICU LOS, respectively), in response to those requests.

RECENT POLICY DECISIONS RELATED TO ICU OUTCOMES

Evaluation of ICU Outcome Risk-Adjustment Models by the National Quality Forum and the Hospital Quality Alliance

During the National Quality Forum comment periods and deliberations, concern was raised about the potential for code status (whether a patient or his family allowed the hospital to provide all possible resuscitative support) would influence performance. In addition, some in the critical care community were concerned that public reporting of ICU performance would create an incentive for referral institutions to refuse to accept complex cases in transfer. Thus, although both the National Quality Forum and the Hospital Quality Alliance endorsed the models, they requested that transfer patients be excluded and a variable be included indicating "full code" status or not. In addition, both organizations stipulate that any performance measure cannot be proprietary, so adopting versions of these models that were copyrighted was not an option.

Response to National Quality Forum and Hospital Quality Alliance Evaluation of ICU Outcome Risk Adjustment Models

At the time of National Quality Forum and the Hospital Quality Alliance endorsement, no existing ICU outcomes model had been calibrated and validated to meet these specifications. We therefore started with the MPM_0-II model (the last non-proprietary version of this model) and added the full code variable, then assessed the model on a population of that excluded transfers. Because of concerns about calibration without considering interactions among the clinical variables,[25] we convened a clinical panel to suggest candidate interactions to be evaluated.

HOSPITAL AND PATIENT SAMPLE

Hospital Sample

The participating hospitals were those who voluntarily contribute patient-level data for public reporting of ICU outcomes in the state of California. In 2009, this sample consisted of 196 hospitals, representing a diverse group of institutions. Hospitals were asked to collect data on the first 100 consecutive patients per quarter who were discharged from the hospitals and had a stay in any of the hospitals' ICUs.

Patient Sample

The inclusion and exclusion criteria reflect the parameters already established by the pre-existing Mortality Probability Model (MPM_0-II), from which the $ICOM_{mort}$ and $ICOM_{LOS}$ models evolved. In 2009, there were 68,122 eligible patients with complete data for risk adjusted mortality calculation. This sample was split into a 60% development set (40,395) and 40% validation set (27,187) for analyses.

Inclusion criteria:

1. Age 18 or older

The original model was developed on a population ≥ 18 years old.[26] The clinical spectrum of diseases for children is significantly different from adult illnesses.

2. Stay in the ICU for at least four hours

Patients are sometimes admitted to the ICU for very short stays for a variety of administrative reasons (such as absence of other beds) or for periprocedural sedation. If these are the only reasons for an ICU stay, these patients are quite distinct from the typical ICU population. Another group of patients with short ICU stays—those who die within a few hours of admission—likely have outcomes determined entirely by clinical events occurring—and care provided prior to—the ICU stay. Therefore, all patients with ICU stays less the four hours are excluded.

Exclusion criteria:

1. Burn patients

Burn patients were excluded from the original model's development population. 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.[31]

2. Trauma patients

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 while other hospitals in the region would not. Furthermore, specific prognostic systems have been previously developed for trauma patients[32] and would be more useful for assessing the performance of regional trauma centers (if this were 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. Like trauma and burn patients, specific prognostic systems have been previously developed for this subset of patients as well.[33]

4. Patients admitted to rule out myocardial infarction (MI) that are found within 24 hours of ICU admission to not have a MI or another critical illness

Individuals who "rule out" for MI essentially are admitted to the ICU for monitoring of chest pain or a similar symptom. When this symptom is not due to 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, rule out MI patients who are found not have an MI or other critical illness are excluded.

5. Readmissions

Readmissions to the ICU during the same stay are excluded since interventions during the first ICU admission may impact the patient's risk of mortality in the second admission.

6. Transfers from another acute care hospital

Pre-ICU treatments have the potential to alter the relationship between physiologic scoring and outcome.[34] The relationship between lead-time bias and patient outcomes is complex, having inconsistent effects on outcome, and often differing by patient type.[25] Previous reports have also demonstrated the potential negative impact that patients transferred into ICUs might have on the accepting center's outcome measures.[35,36] Transferred patients are therefore excluded from the sample.

MODEL DESCRIPTION AND DEVELOPMENT

The $ICOM_{mort}$ and $ICOM_{LOS}$ models evolved from the MPM_0-II, which itself was developed as an updated and revised version of the original MPM. The goal of the MPM 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.[26] 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 primary reason for admission. All variables were collected in the window from one hour prior to ICU admission to one hour after ICU admission. Link to Data dictionary. Link to Data Collection tool.

Assessment of Interactions among Clinical Risk Factors

In our dataset, we found that a base model containing only the MPM_0-II variables with the addition of a full code status variable after excluding transferred patients overpredicted mortality, particularly in the higher ranges of risk. Given this overprediction, the most plausible explanation was that there were interactions among clinical variables for those patients with multiple risk factors. For this reason, we convened a clinical panel that suggested evaluation of the following interactions (Table 1).

Table 1. Candidate Interactions between Selected ICOM Clinical Variables

Interaction terms	Rationale for assessing possible interaction
Acute renal failure x chronic renal insufficiency	In the absence of chronic renal insufficiency, a greater insult is required to cause acute renal failure.
Acute renal failure x systolic blood pressure ≤ 90	Low systolic blood pressure may cause acute renal failure.
GI bleed x heart rate ≥ 150 beats/min	High heart rate may indicate a worse GI bleed.
GI bleed x systolic blood pressure ≤ 90	Low systolic blood pressure may indicate a worse GI bleed.
GI bleed x cirrhosis	GI bleed likely to be worse with cirrhosis, but otherwise protective.
CPR before admission x mechanical ventilation	Representative of combined signs of severe insult.
CPR before admission x	Representative of combined signs of severe insult.

coma/deep stupor	
Coma/deep stupor x mechanical ventilation	Representative of combined signs of severe insult
Cerebrovascular incident x coma/deep stupor	Representative of more severe cerebrovascular insult.
Cerebrovascular incident x intracranial mass effect	Representative of more severe cerebrovascular insult.
Intracranial mass effect x coma/deep stupor	Representative of more severe intracranial mass effect.
Cardiac dysrhythmia x heart rate ≥ 150 beats/min	Dysrhythmia may cause higher heart rate.

Treatment of Age in $ICOM_{mort}$

Prior research has shown that the relationship between age and risk is not necessarily simply linear.[25] In our development dataset, univariate plots of mortality risk versus age suggested increasing risk starting in the mid 60s (most particularly at age 65) and again in the mid 80s (most particularly at age 84). Therefore, we modeled age using splines with knots at 65 and 84. These were implemented by including a term which is the maximum of 0 or age minus the knot value.

The complex relationship of age with the other MPM risk factors was further evaluated using age interaction terms similar to the methods used by the developers of the MPM_0-III.[25] Interactions were considered between age and all of the other MPM risk factors.

Length of Stay Model

For the $ICOM_{LOS}$ model, we used methods similar to those in which we previously validated a model using the same set of variables on patients from 2001-2004.[17] For this model, LOS was truncated at 30 days. Variables and candidate interactions were the same as those evaluated in $ICOM_{mort}$.

Estimation of Models

In other studies in which the 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 has shown poor calibration.[20-24] To improve the calibration of our model, we re-estimated the coefficients in the models on our local sample using methods similar to prior studies that also customized the models to new populations. Therefore we divided our data into a randomly selected model development set (60% of the sample, the group on which the model variables were selected) and a model validation set (40% of the sample, the group on which we confirmed adequate calibration).

RESULTS

The coefficients of our customized models for the estimation samples are shown in Table 2 below.

Table 2. $ICOM_{mort}$ and $ICOM_{LOS}$ Model Re-estimated Coefficients

Variable	$ICOM_{mort}$		$ICOM_{LOS}$	
	Coefficient	p-value	Coefficient	p-value
Constant	-5.707	<.0001	0.032	0.9491
Physiology				
Coma/deep stupor (GCS 3 or 4)	1.037	0.0017	1.871	0.0003

Heart rate ≥ 150 beats/min	2.020	<.0001	1.347	0.0105
Systolic blood pressure ≤ 90	0.919	<.0001	1.257	<.0001
Chronic diagnoses				
Chronic renal insufficiency	0.939	0.0002	0.267	0.4096
Cirrhosis	1.693	0.0015	0.827	0.2802
Metastatic neoplasm	2.826	<.0001	0.993	0.0082
Acute diagnoses				
Acute renal failure	1.588	<.0001	2.056	<.0001
Cardiac dysrhythmia	-0.181	0.4104	0.305	0.2424
Cerebrovascular incident	1.655	<.0001	1.963	<.0001
GI bleed	0.536	0.1206	-0.835	0.0305
Intracranial mass effect	-0.171	0.7102	0.549	0.3269
Other				
Age (per year)	0.032	<.0001	0.017	0.0173
Age spline age 65	0.011	0.0177	-0.008	0.1590
Age spline age 84	0.022	0.0268	-0.032	0.0346
CPR before admission	1.766	<.0001	-0.334	0.5464
Mechanical ventilation within 1 hr of admission	1.388	<.0001	2.738	<.0001
Medical or unscheduled surgical admit	2.404	<.0001	1.630	<.0001
Zero factors (no factors other than age from list above)	-0.034	0.9746	1.281	0.0040
Full code	-1.691	<.0001	0.183	0.6154
Interaction terms between clinical variables				
Acute renal failure x chronic renal insufficiency	-0.615	<.0001	-0.382	0.0130
Acute renal failure x systolic blood pressure ≤ 90	-0.203	0.0119	-0.503	0.0001
GI bleed x heart rate ≥ 150 beats/min	-0.345	0.4228	0.479	0.4882
GI bleed x systolic blood pressure ≤ 90	0.126	0.3542	-0.575	0.0047
GI bleed x cirrhosis	-0.389	0.0834	0.064	0.8435
CPR before admission x mechanical ventilation	0.281	0.1791	0.299	0.3617
CPR before admission x coma/deep stupor	0.259	0.1249	-0.440	0.1486
Coma/deep stupor x mechanical ventilation	-0.545	0.0023	-1.030	0.0009
Cerebrovascular incident x coma/deep stupor	0.209	0.3219	-0.201	0.5927
Cerebrovascular incident x intracranial mass effect	0.784	0.0002	0.278	0.3328
Intracranial mass effect x coma/deep stupor	0.950	0.0003	-1.731	0.0001
Cardiac dysrhythmia x heart rate ≥ 150 beats/min	-0.535	0.0113	-0.626	0.0602
Interaction terms between age and other clinical variables				
Age x coma/deep stupor	0.002	0.5944	-0.012	0.0908
Age x heart rate ≥ 150 beats/min	-0.015	0.0048	-0.010	0.2251
Age x systolic blood pressure ≤ 90	0.000	0.8927	-0.007	0.0337
Age x chronic renal insufficiency	-0.005	0.1427	0.002	0.6226
Age x cirrhosis	-0.009	0.2827	-0.013	0.2888
Age x metastatic neoplasm	-0.023	<.0001	-0.008	0.1678
Age x acute renal failure	-0.010	0.0001	-0.014	0.0001
Age x cardiac dysrhythmia	0.003	0.2390	-0.001	0.7962
Age x cerebrovascular incident	-0.017	0.0004	-0.019	0.0021
Age x GI bleed	-0.009	0.0593	0.013	0.0163
Age x intracranial mass effect	0.002	0.7830	-0.001	0.8684

Age x CPR before admission	-0.013	0.0010	-0.004	0.5739
Age x mechanical ventilation	-0.004	0.0761	-0.005	0.0734
Age x medical or unscheduled surgical admit	-0.018	0.0006	-0.012	0.0122
Age x zero factors	-0.005	0.7081	-0.017	0.0107
Age x full code	0.011	0.0005	0.007	0.1300

$ICOM_{mort}$ Model Performance

Discrimination was assessed by using the area under the receiver operating characteristic curve (AUC). The minimum AUC that was considered reasonable discrimination was 0.80.[37] Our model demonstrated adequate discrimination on the validation sample, with an AUC of 0.820 (Table 3).

Calibration was assessed using the Hosmer-Lemeshow goodness-of-fit tests and calibration curves. We performed both the Hosmer-Lemeshow C test and H test. 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 the chi-square distribution. The degrees of freedom equal N-2, where N= 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.)[37]

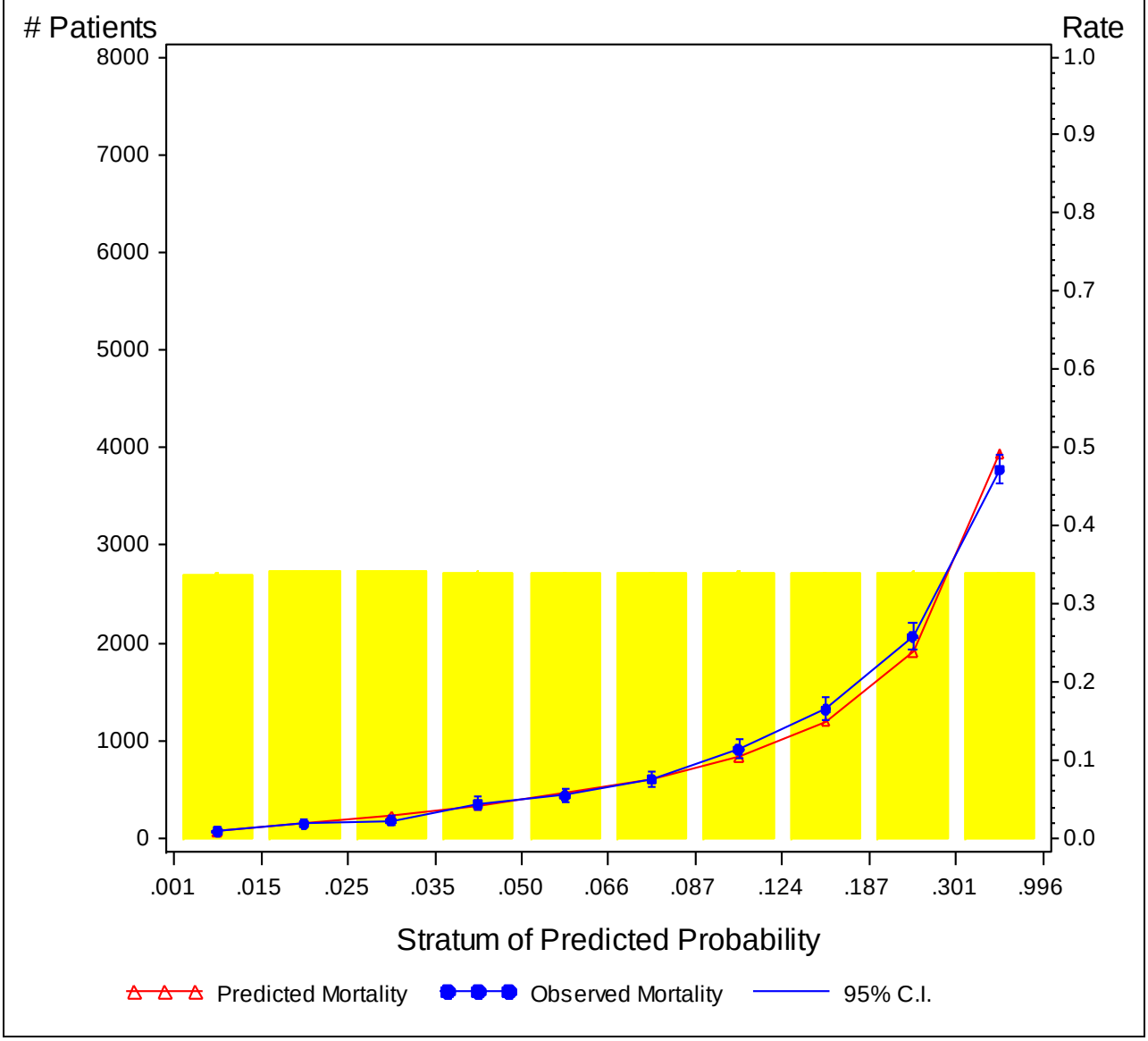
Given the sensitivity of the Hosmer-Lemeshow statistic to sample size,[37,38] after recalibration of the coefficients using logistic regression, calibration was reassessed using 11 random samples of 5,000 patients (Table 3) taken from the validation sample.[38] Nine of the 11 randomly selected samples of 5,000 patients showed non-significant H-L statistics. Calibration was also assessed using the adjunct measure of a graph plotting observed vs. predicted mortality. This plot is depicted in Figure 1.

Table 3. Performance of the Re-estimated $ICOM_{mort}$ Model on the Validation Sample

AUC (95% CI)	**Median H-L statistic***	
	C-test (p-value)	**H-test (p-value)**
0.820 (0.813-0.828)	12.04 (0.28)	16.85 (0.08)

*H-L statistics calculated on 11 random samples of 5,000 patients from the validation sample.

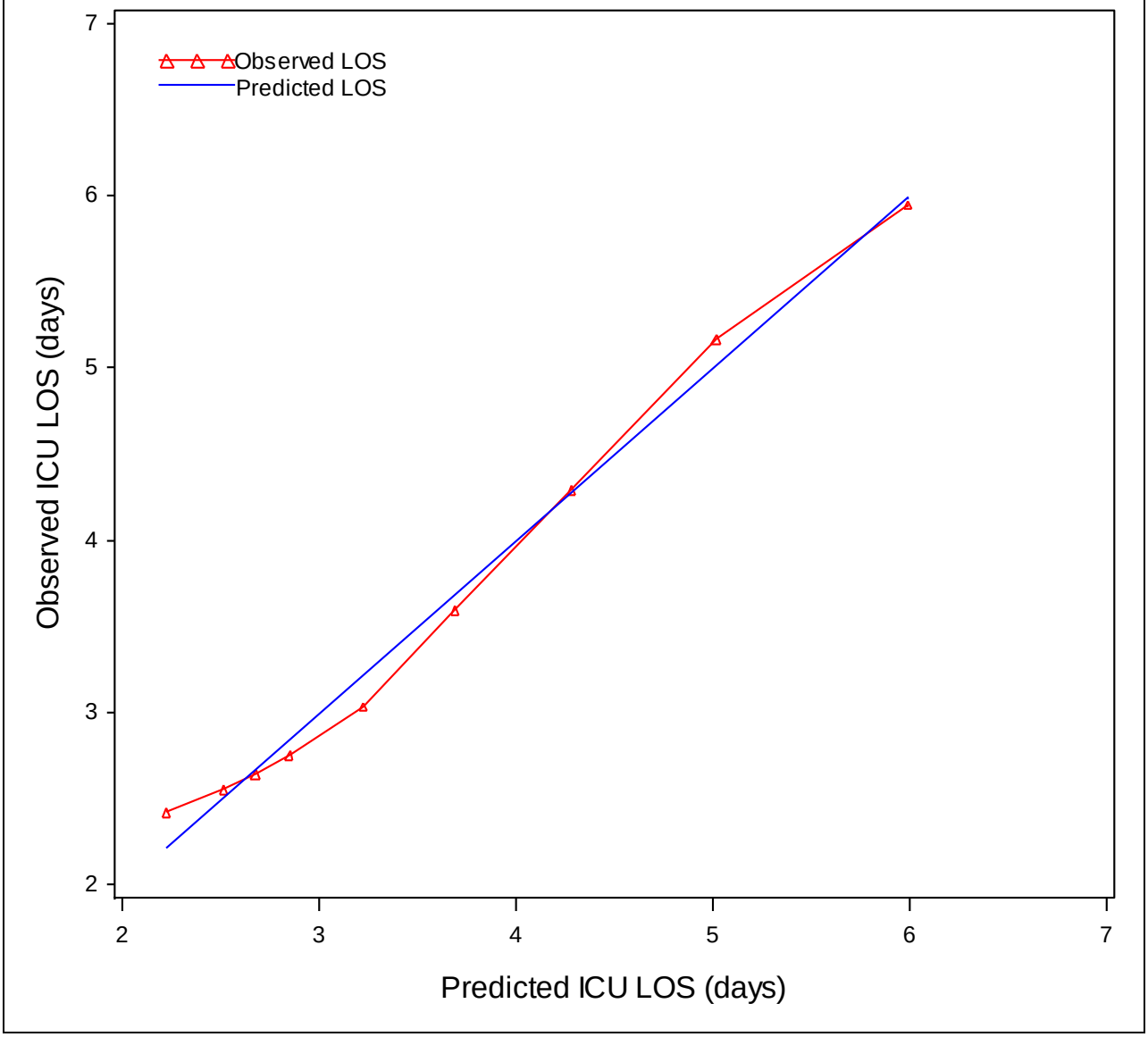
Figure 1. Calibration Curve for ICU Outcomes for Mortality ($ICOM_{mort}$)



$ICOM_{LOS}$ Model Performance

The $ICOM_{LOS}$ model was estimated using linear regression, with the randomly selected 60% estimation sample. The model showed moderate predictive power with adjusted R-square of 0.082. The calibration of the model was assessed within the 40% validation sample. Within deciles of predicted LOS, the following plot (Figure 2) compares the predicted versus observed mean LOS.

Figure 2. Calibration Curve for ICU Outcomes for Length-of-Stay ($ICOM_{LOS}$)



Because the patients used in evaluating LOS included some patients who died (approximately 12%), some elements of the model predicting LOS reflect the fact that death is associated with shorter than expected lengths of stay. This is reflected in the fact that the signs on some of the coefficients of some variables in $ICOM_{LOS}$ are the opposite of those in $ICOM_{mort}$ (that is, some variables with negative coefficients in $ICOM_{LOS}$ have positive coefficients in $ICOM_{mort}$ and vice versa), as seen in Table 2.

RECOMMENDATIONS FOR IMPLEMENTATION

It is now widely understood that any risk-adjusted ICU outcomes model needs to be re-calibrated (or even re-estimated) when applied to any new population,[20-24] and the $ICOM_{mort}$ and $ICOM_{LOS}$ are not exceptions to this rule. Therefore, although our best current models would involve using only the statistically significant variables (including interaction terms) from Table 2 for $ICOM_{mort}$ and $ICOM_{LOS}$, these models are not likely to be optimal for long.

Rather, we recommend that individuals or organizations wishing to assess risk-adjusted ICU outcomes collect all the variables required for these models and then recalibrate (or re-estimate) the models to fit the population whose outcomes are being evaluated. As above, we believe it is best to perform these tasks first on a randomly selected model development subsample of the overall population and test discrimination and calibration in a model validation sample. Of note, since the Hosmer-Lemeshow statistic is very sensitive to sample size, we recommend assessing calibration among multiple sets of 5,000 randomly selected members of the validation subsample.[38]

SUMMARY

The new $ICOM_{mort}$ and $ICOM_{LOS}$ models both demonstrate adequate performance as measured by discrimination and calibration. With the lowest data burden of any existing ICU outcomes models, $ICOM_{mort}$ and $ICOM_{LOS}$ provide a standardized means by which hospital and ICU performance can be compared between institutions at a reasonable cost. Prior to application to any new population, however, we recommend re-estimation of coefficients on the local sample in addition to confirmation of relevant interactions.

1. 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.

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

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. 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.

5. 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.

6. 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.

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

8. Halpern NA, Pastores SM. Critical care medicine in the United States 2000-2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med.* Jan 2010;38(1):65-71.

9. Rapoport J, Teres D, Lemeshow S, Avrunin JS, Haber R. Explaining variability of cost using a severity-of-illness measure for ICU patients. *Med Care.* Apr 1990;28(4):338-348.

10. Rapoport J, Teres D, Lemeshow S, Gehlbach S. A method for assessing the clinical performance and cost-effectiveness of intensive care units: a multicenter inception cohort study. *Crit Care Med.* Sep 1994;22(9):1385-1391.

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

12. 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.

13. Kuzniewicz MW, Vasilevskis EE, Lane R, et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. *Chest.* Jun 2008;133(6):1319-1327.

14. Rothen HU, Stricker K, Einfalt J, et al. Variability in outcome and resource use in intensive care units. *Intensive Care Med.* Aug 2007;33(8):1329-1336.
15. Knaus WA, Wagner DP, Zimmerman JE, Draper EA. Variations in mortality and length of stay in intensive care units. *Ann Intern Med.* May 15 1993;118(10):753-761.
16. Render ML, Kim HM, Deddens J, et al. Variation in outcomes in Veterans Affairs intensive care units with a computerized severity measure. *Crit Care Med.* May 2005;33(5):930-939.
17. Vasilevskis EE, Kuzniewicz MW, Cason BA, et al. Mortality probability model III and simplified acute physiology score II: assessing their value in predicting length of stay and comparison to APACHE IV. *Chest.* Jul 2009;136(1):89-101.
18. Rosenthal GE, Harper DL, Quinn LM, Cooper GS. Severity-adjusted mortality and length of stay in teaching and nonteaching hospitals. Results of a regional study. *JAMA.* Aug 13 1997;278(6):485-490.
19. Woods AW, MacKirdy FN, Livingston BM, Norrie J, Howie JC. Evaluation of predicted and actual length of stay in 22 Scottish intensive care units using the APACHE III system. Acute Physiology and Chronic Health Evaluation. *Anaesthesia.* Nov 2000;55(11):1058-1065.
20. 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.* Sep 2002;30(9):1995-2002.
21. 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(1):31-36.
22. 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.* Dec 1996;24(12):1968-1973.
23. 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.
24. 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.* Jan 1996;24(1):57-63.
25. 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.* Mar 2007;35(3):827-835.
26. 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.
27. 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.
28. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med.* May 2006;34(5):1297-1310.
29. 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.
30. 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.* Oct 2005;31(10):1345-1355.
31. Galeiras R, Lorente JA, Pertega S, et al. A model for predicting mortality among critically ill burn victims. *Burns.* Mar 2009;35(2):201-209.

32. Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. *J Trauma.* Apr 1987;27(4):370-378.
33. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg.* Jul 1999;16(1):9-13.
34. Tunnell RD, Millar BW, Smith GB. The effect of lead time bias on severity of illness scoring, mortality prediction and standardised mortality ratio in intensive care--a pilot study. *Anaesthesia.* Nov 1998;53(11):1045-1053.
35. Combes A, Luyt CE, Trouillet JL, Chastre J, Gibert C. Adverse effect on a referral intensive care unit's performance of accepting patients transferred from another intensive care unit. *Crit Care Med.* Apr 2005;33(4):705-710.
36. Rosenberg AL, Hofer TP, Strachan C, Watts CM, Hayward RA. Accepting critically ill transfer patients: adverse effect on a referral center's outcome and benchmark measures. *Ann Intern Med.* Jun 3 2003;138(11):882-890.
37. Mourouga P, Goldfrad C, Rowan KM. Does it fit? Is it good? Assessment of scoring systems. *Current Opinion in Critical Care* June 2000;6(3):176-180.
38. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med.* Sep 2007;35(9):2052-2056.