Performance of APACHE III Models in an Australian ICU*

David A. Cook, B Med Sci, MB BS

Study objective: Evaluation of the performance of the APACHE (acute physiology and chronic health evaluation) III ICU and hospital mortality models at an Australian tertiary adult ICU.

Design: Noninterventional, observational study.

Setting: Metropolitan, Australian, tertiary referral medical/surgical ICU.


Measurements: Prospective collection of demographic, diagnostic, physiologic, laboratory, admission, and discharge data.

Results: The patient sample was younger and more commonly male, with more comorbidities and a different operative and referral source mix, compared to the APACHE III development sample. Receiver operating characteristic curve areas for ICU (0.92) and hospital mortality (0.90) demonstrated excellent discrimination. Observed ICU mortality (9.9%) did not significantly differ from the prediction of the APACHE III model (8.9%) or the APACHE III model adjusted for hospital characteristics (10.5%). The hospital mortality (16.0%) was underestimated by the APACHE III model [13.6%; $\chi^2(1) = 7.4; p = 0.01$]. With proprietary adjustments for hospital characteristics (14.9%) or referenced to the US database (15.6%), agreement was closer. Good calibration was found with all models except the unadjusted hospital mortality model.

Conclusion: In contrast to other non-American studies, this Australian study demonstrates that the APACHE III can perform well on independent assessment. As perfect discrimination and calibration cannot coexist in a probabilistic model with dichotomous outcomes, performance of APACHE III models with proprietary adjustment for hospital characteristic provide a good compromise for use in quality surveillance.

Key words: acute physiology and chronic health evaluation; Australia; critical care; hospital mortality; intensive care; outcome prediction; severity of illness

Abbreviations: APACHE = acute physiology and chronic health evaluation; H-L = Hosmer-Lemeshow; PAH = Princess Alexandra Hospital; ROC = receiver operating characteristic

Princess Alexandra Hospital (PAH) ICU provides medical and surgical critical-care services to an 858-bed adult metropolitan hospital that is the regional center for trauma, major surgery, medical subspecialties, and psychiatry. In August 1994, the APACHE (acute physiology and chronic health evaluation) III management system1 was introduced to the 12-bed ICU.

The APACHE III mortality estimates are part of a proprietary database and decision support system provided by Apache Medical Systems (McLean, VA). The risk equations and weights were developed by Knaus and colleagues.2 The APACHE III score is an ICU severity of illness score calculated from the patient’s age, the presence of comorbid conditions, and the worst physiologic and laboratory investigations in the first 24 h. The APACHE III risk estimate equations use the admission diagnosis, the source of admission, and the APACHE III score weighted according to coefficients that are not in the public domain.

The performance of the APACHE III first-day predictions have been evaluated on the development database.2 Independent APACHE III validation series are available from Brazil (multicenter, 1,734 patients),3 the United Kingdom (single institution, 1,144 patients;4 multicenter, 12,793 patients), and Germany (single institution, 2,661 patients).6 In each study, hospital mortality was higher than predicted with resultant poor model calibration. In contrast, in two large, prospective, multicenter North American
series (37,668 patients and 116,340 patients), the APACHE III model demonstrated good overall performance.

The APACHE III mortality prediction model has not performed well in clinical evaluation outside the United States, where it was developed. An independent evaluation of neither the ICU mortality model nor of the proprietary-adjusted predictive models was available. The purpose of this study was to independently assess the performance of the APACHE III hospital and ICU mortality models, unadjusted and with proprietary model adjustments, at an Australian ICU.

**Materials and Methods**

Admissions to the PAH ICU were studied from January 1, 1995, to December 31, 1997. Patients <16 years old, cardiac surgical and burns patients, and patients admitted for <4 h or for exclusion of myocardial infarction were excluded from the APACHE III predictions. Patient data were collected according to the rules of Apache III. Data were manually collected or transferred from the pathology laboratory information system. The database manager verified all data. After the first 6 months of the data collection, 4% of patient records were extracted to ensure interreporter reliability. Outcome was survival status on discharge from the ICU or the PAH. Patients transferred to rehabilitation facilities (spinal, geriatric, head injury, and general rehabilitation units) or the psychiatric unit within the PAH complex were deemed inpatients until discharged from the campus.

The study was conducted with the approval of the hospital research ethics committee, using de-identified data. Analysis was carried out using Excel 97 and Access 97 software (Microsoft; Redmond, WA) and Statistica 5.1 software (Statsoft; Tulsa, OK). Comparisons were made between the APACHE III developmental patient sample and the current study sample using t test or $\chi^2$ test, adopting a significance level of $p < 0.01$ to correct for multiple testing.

The ICU mortality models were assessed on all eligible admissions to ICU, including readmissions. Hospital mortality model assessment excludes all ICU readmissions during an episode of hospitalization. For each admission, mortality estimates were provided by proprietary weights and the APACHE III equation. For in-ICU mortality, the APACHE III ICU mortality model and a model with proprietary adjustments for hospital characteristics (similar hospital ICU mortality model) were studied. Three models of hospital mortality were evaluated. The APACHE III hospital mortality model and models with proprietary adjustments for hospital characteristics (similar hospital ICU mortality model) were studied. The APACHE III hospital mortality model and models with proprietary adjustments for hospital characteristics (similar hospital ICU mortality model) were studied. The APACHE III hospital mortality model references the predictions to teaching hospitals of similar size in the US Midwest region. The US database hospital mortality model reflects a “typical” US ICU modeled from the APACHE III database (C. Alzola, MS; personal communication; November 1999).

The aggregate predicted mortality rate for each model was the sum of estimated probabilities of death divided by the number of admissions. The standardized mortality ratio was the ratio of observed mortality to the aggregate predicted mortality. Confidence intervals were estimated using a normal approximation to the binomial distribution.

For assessment of model fit or calibration, the agreement between predicted and observed mortality rate in risk ranges was assessed. Calibration curves (Fig 1, 2) using 10 equal, contiguous risk ranges present observed against predicted outcomes with 95% confidence intervals estimated by continuity corrected normal approximations to binomial distributions. The Hosmer-Lemeshow (H-L) statistics indicate the agreement between the observed and predicted mortality across risk ranges. For C*, admissions are ranked according to predicted risk of death and divided into 10 near-equivalent groups. $H^*$ uses the sample divided into 10 contiguous decimals of risk of equal span, but unequal number. The C* and $H^*$ statistics are like a Pearson $\chi^2$ statistic calculated from a $4 \times 10$ table of observed and estimated mortality and survival. For external validation studies, the degrees of freedom of the $\chi^2$ distribution is the number of ranges of risk.
Rejection of the null hypothesis that there is no difference between the predicted frequencies across the deciles of risk is at \( p < 0.05 \).

Discrimination was assessed by calculating area under the receiver operating characteristic (ROC) curves, with estimates of the standard error and confidence intervals.\(^\text{13}\) The area under the ROC curve estimates the probability that a randomly selected mortality will be given a higher risk of death estimate than a randomly chosen survivor. It is a global measure of the ability of the model to assign a higher risk of death to patients who die.\(^\text{14}\)

**Results**

There were 3,455 admissions to the PAH ICU between January 1, 1995, and December 31, 1997. Exclusions were 45 admitted patients < 16 old, 8 staying < 4 h, and 4 with burns. The 3,398 remaining admitted patients represented 3,159 patient hospitalizations of 3,038 individual patients. All patient outcomes were accounted for during the study period.

There were 338 deaths in ICU (9.9%) and 507 deaths in hospital (16.0%). The median length of stay in ICU was 2 days (range, 1 to 75 days), with 65.4% of patients admitted for 2 or 3 days. Median duration of hospitalization was 16 days (range, 1 to 930 days; 25% quartile, 8 days; 75% quartile, 28 days).

Compared to the APACHE III development sample, the PAH sample was younger, had a greater male preponderance, had a different case mix of nonoperative/operative patients (elective and emergency), and a different mix of sources of referral. Severity of illness, reflected by the day 1 APACHE III score and the acute physiology score component appear similar (Table 1).

The 231 admission diagnoses were mapped onto 77 disease groups. The commonest operative disease groups were GI cancer (9.0%), elective aortic surgery (8.5%), operative trauma (7.2%), head and neck cancer surgery (3.5%), miscellaneous GI surgery (3.5%), and liver transplantation (2.8%). The commonest nonoperative groups were nonoperative trauma (10.3%), drug overdose (7.5%), cardiac arrest (2.6%), and asthma (2.5%). The 10 most frequent groups accounted for 57.3% of all admissions.

There were 2,812 admissions (82.8%) with no APACHE III comorbidities; 459 admissions (13.5%) had one comorbidity, 120 admissions (3.5%) had two comorbidities, and 7 admissions (0.02%) had three or more comorbidities. The prevalence of one or more comorbidities in the present study sample (17.2%) differs from that of the APACHE III developmental group of 6.6% \( (\chi^2, 1,735; \text{degrees of freedom}, 1; p < 0.0001) \).

Figure 2. Calibration curves for APACHE III hospital mortality models, showing calibration curves of observed mortality with 95% confidence intervals against predicted mortality. Patients admitted to PAH ICU are grouped in decile risk ranges of APACHE III hospital mortality estimates. The dashed line is that of perfect agreement between observed and predicted mortalities. Top: Hospital mortality model. Middle: Similar hospital mortality model, with proprietary adjustment for hospital characteristics. Bottom: US database hospital mortality model, with proprietary adjustment to reflect units on the US database.
Patients had a predicted ICU mortality of ≤ 0.1, and 91% had ≤ 0.3. Sixty-eight percent of patients had a predicted hospital mortality of ≤ 0.1, and 85% had ≤ 0.3.

The observed hospital mortality (16.0%) was significantly higher than the APACHE III predicted mortality rate of 13.6% \( \chi^2(1), 7.4; p = 0.01 \) [Table 2]. The observed hospital mortality rate was not different to the APACHE III predictions when model adjustments are made for hospital characteristics (similar hospital model, 14.9%) or when the US database-referenced model (15.6%) is used. The observed ICU mortality rate (9.9%) was not significantly different from the predictions of the APACHE III ICU mortality model (8.9%) or the APACHE III similar hospital ICU model (10.5%; Table 3).

The APACHE III ICU models show good calibration. Calibration curves for the ICU mortality model and the similar hospital ICU mortality model (Fig 1), have the line of perfect model agreement lying within the 95% confidence intervals for all risk ranges. The H-L statistics (Table 4), with corresponding \( p \) values > 0.05, confirm adequate calibration, with the similar hospital ICU model providing the best fit on this sample.

The similar hospital model and the model referenced to the US database display the adequate calibration. Although both calibration curves show that observed mortality differs from expected in the range of 40 to 50% (Fig 2, middle, bottom), the H-L statistics (Table 4) infer nonsignificant deviations from perfect fit. The unadjusted APACHE III hospital mortality model has poor calibration demonstrated on the calibration curve (Fig 2, top) and on statistical analysis (Table 4).

The area under the ROC curves for both ICU mortality models was 0.92. The areas under the ROC

### Table 1—Comparison of Demographics, Operative Status, and APACHE III Score Between Princess Alexandra Hospital ICU Admissions and APACHE III Developmental Sample*

<table>
<thead>
<tr>
<th>Variables</th>
<th>PAH ICU</th>
<th>APACHE Developmental Sample†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size (ICU admissions), No.</td>
<td>3,159</td>
<td>17,440</td>
</tr>
<tr>
<td>Age, yr (SD)†</td>
<td>52.6 (19.2)</td>
<td>59 (19)</td>
</tr>
<tr>
<td>Male‡</td>
<td>61.5</td>
<td>55.2</td>
</tr>
<tr>
<td>Operative status¶</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonoperative</td>
<td>46.8</td>
<td>57.7</td>
</tr>
<tr>
<td>Elective</td>
<td>38.4</td>
<td>33.3</td>
</tr>
<tr>
<td>Emergency</td>
<td>14.9</td>
<td>9.0</td>
</tr>
<tr>
<td>Admission source</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operating/recovery department</td>
<td>53.2</td>
<td>42.3</td>
</tr>
<tr>
<td>Emergency department</td>
<td>20.0</td>
<td>35.5</td>
</tr>
<tr>
<td>Hospital ward</td>
<td>13.8</td>
<td>16.4</td>
</tr>
<tr>
<td>Other hospital/ICU</td>
<td>13.0</td>
<td>5.8</td>
</tr>
<tr>
<td>Day 1 APACHE III score, mean</td>
<td>47.6</td>
<td>49.2</td>
</tr>
<tr>
<td>Day 1 acute physiology score, mean</td>
<td>39.2</td>
<td>39</td>
</tr>
<tr>
<td>ICU mortality</td>
<td>10.0</td>
<td>NA</td>
</tr>
<tr>
<td>Hospital mortality</td>
<td>16.0</td>
<td>16.8</td>
</tr>
</tbody>
</table>

*Data are presented as % unless otherwise indicated. NA = information not available from publications.
†Data from Knaus et al² and Wagner et al.35
‡Two-tailed \( t \) test, \( p < 0.001 \).
¶\( \chi^2(1), 46; p < 0.001 \).
¶¶\( \chi^2(2), 180; p < 0.001 \).
||\( \chi^6(6), 515; p < 0.001 \).

### Table 2—Predicted APACHE III Hospital Mortality Compared to Observed Hospital Mortality Rate of 16.0%*

<table>
<thead>
<tr>
<th>Variables</th>
<th>Predicted Mortality Rate, %</th>
<th>SMR (95% CI)</th>
<th>( \chi^2(1) )</th>
<th>( p ) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospital mortality model</td>
<td>13.6</td>
<td>1.17 (1.08–1.29)</td>
<td>7.4</td>
<td>0.01†</td>
</tr>
<tr>
<td>Similar hospital mortality model</td>
<td>14.9</td>
<td>1.07 (1.00–1.17)</td>
<td>1.8</td>
<td>0.19</td>
</tr>
<tr>
<td>US database hospital mortality model</td>
<td>15.6</td>
<td>1.02 (0.96–1.12)</td>
<td>0.3</td>
<td>0.63</td>
</tr>
</tbody>
</table>

*SMR = standardized mortality ratio (observed/predicted mortality); CI = confidence interval.
†Significant at \( p < 0.05 \) with no adjustment for multiple testing.

CHEST / 118 / 6 / DECEMBER, 2000
curves for the APACHE III hospital mortality model and the model referenced for US database were 0.90. For the similar hospital model, the area was 0.91.

**Discussion**

This study is the largest single institution prospective assessment of APACHE III outside the United States and the first study to evaluate the performance of the APACHE III ICU and hospital mortality models with proprietary adjustments for hospital characteristics and referenced to the US database. It demonstrates that the APACHE III mortality models with adjustment for hospital characteristics, the ICU mortality model, and the hospital mortality model referenced to the USA database have good discrimination and calibration in an Australian adult ICU population. This is the first series from a general ICU outside the United States that endorses APACHE III model performance. It also supports the findings of previous reports from the United Kingdom,4,5 Brazil,3 and Germany,6 where the original, unadjusted hospital mortality model2 performed poorly.

There is controversy about the best assessment of calibration of probabilistic models of dichotomous outcomes. The reader is directed to reviews of the methodologies.14–18 The practical approach taken in this study is the graphical presentation of calibration curves and testing of goodness-of-fit with the H-L statistics. This allows qualitative graphical comparison with other published calibration curves for APACHE III3–5,7 and other models. The calibration at the PAH resembles the curves of the North American ICUs7 and the Brazilian ICUs3 selected for post hoc analysis on the basis of low standardized mortality ratios.

Conclusions from the H-L statistics should be drawn with consideration of limitations of the method and in light of other relevant evidence. As $\chi^2$-like statistics, the magnitude of the H-L statistics are not only dependent on the fit or calibration of the model, but on absolute patient numbers and the distribution of the estimates.17 These methods were developed for comparison of models on a common data set,11 where patient numbers and distribution are controlled. An application of prospective independent validation using $\chi^2$ distributions with 10 degrees of freedom12,16 is used in this study. If the analysis is repeated using eight degrees of freedom, as adopted by some authors, the p values are correspondingly smaller. In this sample, the H-L statistics, the calibration curves, and the global agreement between observed and predicted outcomes concur that among the models studied, only the unadjusted APACHE III hospital mortality model had inadequate calibration.

Discrimination of all APACHE III models was good ROC curve area (> 0.8) or excellent ROC curve area (> 0.9),19 approaching the practical limit of the generalization performance of this type of multivariate model.20 Comparison between the discrimination of APACHE III hospital mortality predictions at PAH and other published series (Table 5) shows similar performance to the developmental set,2 the US prospective multicenter validation series,7,8 and the United Kingdom multicenter series.5 However, the discrimination of the APACHE III model is not invulnerable to all case-mix, clinical

<table>
<thead>
<tr>
<th>Variables</th>
<th>Predicted Mortality Rate, %</th>
<th>SMR (95% CI)</th>
<th>$\chi^2$(1)</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU mortality model</td>
<td>8.9</td>
<td>1.12 (1.01–1.26)</td>
<td>2.5</td>
<td>0.12</td>
</tr>
<tr>
<td>Similar hospital ICU mortality model</td>
<td>10.5</td>
<td>0.95 (0.87–1.06)</td>
<td>0.4</td>
<td>0.32</td>
</tr>
</tbody>
</table>

*See Table 2 for abbreviations.

<table>
<thead>
<tr>
<th>Variables</th>
<th>$C^*$</th>
<th>p Value</th>
<th>$H^*$</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU mortality model</td>
<td>12.7</td>
<td>0.24</td>
<td>14.8</td>
<td>0.13</td>
</tr>
<tr>
<td>Similar hospital ICU mortality model</td>
<td>5.6</td>
<td>0.85</td>
<td>8.7</td>
<td>0.56</td>
</tr>
<tr>
<td>Hospital mortality model</td>
<td>32.8</td>
<td>&lt; 0.001</td>
<td>35.6</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Similar hospital mortality model</td>
<td>14.5</td>
<td>0.15</td>
<td>16.9</td>
<td>0.08</td>
</tr>
<tr>
<td>US database mortality model</td>
<td>12.2</td>
<td>0.27</td>
<td>15.9</td>
<td>0.10</td>
</tr>
</tbody>
</table>

*H-L goodness-of-fit statistic ($C^*$ and $H^*$) with probabilities using $\chi^2$ distribution with 10 degrees of freedom.
performance... “a n dt h e difficulty in separating universally recognized and verified marker of ICU model performance from clinical performance. Dif-

<table>
<thead>
<tr>
<th>Source</th>
<th>Area Under ROC Curve</th>
<th>95% Confidence Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAH</td>
<td>0.90</td>
<td>0.89–0.91</td>
</tr>
<tr>
<td>Knau et al, 1991</td>
<td>0.90</td>
<td>0.89–0.91</td>
</tr>
<tr>
<td>Bastos et al, 1996</td>
<td>0.82</td>
<td>0.80–0.84</td>
</tr>
<tr>
<td>Beck et al, 1997</td>
<td>0.85</td>
<td>0.82–0.88</td>
</tr>
<tr>
<td>Zimmerman et al, 1998</td>
<td>0.89</td>
<td>0.88–0.90</td>
</tr>
<tr>
<td>Pappachan et al, 1999</td>
<td>0.89</td>
<td>Interval &lt; 0.01</td>
</tr>
<tr>
<td>Sirio et al, 1999</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Markgraf et al, 2000</td>
<td>0.85</td>
<td>0.81–0.88</td>
</tr>
</tbody>
</table>

Table 5—Comparison of Discrimination of APACHE III Hospital Mortality Predictions

practice, and data collection conditions, given the good, but lesser performance in the multicenter Brazilian series,3 and single institution studies from England4 and Germany.6

Discrimination of the APACHE III system compares favorably with other first-day hospital mortality predictions for ICU patients. The Mortality Probability Model II achieves ROC curve area of 0.74 to 0.84.19,21–23 The Simplified Acute Physiology Score II has a ROC area of 0.82 to 0.85 on prospective validation samples.6,24,25 APACHE II discrimination ranges between 0.79 and 0.89.6,21,25–27 Organ system failure models28,29 perform well on developmental data sets, although prospective evaluation is limited.

There is little with which to compare the performance of the APACHE III predictions for in ICU mortality. A multiple organ dysfunction score mortality prediction achieved an impressive ROC curve area of 0.93 on a small, prospective validation patient sample drawn from the surgical ICU in which it was developed.30 The estimated confidence intervals are broad (95% confidence intervals, 0.86 to 0.99) due to small numbers, and subsequent independent validation of generalized performance has not been published.

Despite case-mix and referral differences in samples between the PAH and the APACHE III developmental sample, the performance of the APACHE III models adjusted for hospital characteristics was good. Other analyses have only assessed the performance of the unadjusted APACHE III hospital mortality models. A UK validation sample5 with a higher average APACHE III score, more comorbidities, and different referral sources found excellent discrimination, but a 25% higher mortality rate with poor calibration and excess mortality rates in all risk ranges. The authors note “... the lack of a universally recognized and verified marker of ICU performance ...” and the difficulty in separating model performance from clinical performance. Dif-

ferences in case mix were proposed as the likely reason for higher than predicted hospital mortality rate. A German study6 describes a 22% higher-than-predicted mortality rate, with data collection anomalies, case mix, lead-time bias, model inaccuracy, or quality of care sited as possible contributors.

Performance of ICU mortality prediction models that do not use admission diagnosis, such as Mortal-

ity Probability Model II, are quite vulnerable to changes in case mix19 and even to mortality rates.22 In contrast to other work,3–6 the present study observes that the APACHE III model can be applied to a patient population with different case mix and referral pattern, outside of the United States, and produce similar performance to that observed on the APACHE III development2 and validation series.7

In a probabilistic outcome prediction model, perfect calibration and perfect discrimination are incompatible.31 A model can achieve perfect discrimi-

nation if it becomes a dichotomous classifier, therefore no longer providing a probabilistic estimate of risk. Furthermore, a model that is perfectly calibrated may have a ROC area as low as 0.83.31 Maximization of model performance is, therefore, a compromise. In the PAH series, the APACHE III mortality estimates, particularly with proprietary ad-

justment for hospital characteristics, provided both good discrimination and good calibration. This supports the validity and robustness of APACHE III variables, data collection, and the mortality estimates model.

The local performance of APACHE III has al-

lowed the application of risk-adjusted industrial quality surveillance methods within this ICU. At the PAH ICU, the database function, standard and ad hoc reporting, APACHE III severity of illness, and mortality risk estimates are part of the unit quality assurance program. Risk-adjusted CUSUM charts32,33 control charts, and exponential mean weighted average charting offer qualitative monitoring of risk-adjusted patient mortality. These tech-

niques have been valuable for tracking the outcome effects of institutional changes in the process of intensive care. In 1998 to 1999, the cost of the APACHE III Management System, including a database manager salary, represented less than 1% of the direct ICU costs of the PAH ICU.

The validation of the generalization performance of a risk prediction tool on independent data sets implies model validity and reliability of variables and data collection methods.30 Potential for bias and inaccuracy34 and threats to model performance can arise from local anomalies of clinical practice, case mix, or data collection. The apparent variability of performance of ICU outcome or risk adjustment models mandate that these models must be closely
ACKNOWLEDGMENTS: The author acknowledges the work of Gail Galbraith, Rod Harford, and all who collect the data, and Dr. Anthony Morton for advice and comment.

REFERENCES