Illness severity scoring for Intensive Care at Middlemore Hospital, New Zealand: past and future

Susan L Mann, Mark R Marshall, Alec Holt, Brendon Woodford, Anthony B Williams

Abstract

Aim The Acute Physiological and Chronic Health Evaluation (APACHE) II score is a popular illness severity scoring system for intensive care units. Scoring systems such as the APACHE II allow researchers and clinicians to quantify patient illness severity with a greater degree of accuracy and precision, which is critical when evaluating practice patterns and outcomes, both within or between intensive care units. The study aims to: assess changes in APACHE II scores and hospital-standardised mortality ratio at our ICU over a nine year period from 1 January 1997 to 31 December 2005; assess for changes in the performance of the APACHE II scoring system in predicting patient hospital mortality over the same period; and assess for any clinical subgroups in which APACHE II scoring was particularly inaccurate or imprecise.

Methods Retrospective audit of a single centre relational database, with evaluation of the APACHE II scoring system by year through discrimination (ability to discriminate between the patients who will die or survive at hospital discharge) using receiver operating characteristic (ROC) curves, and calibration (ability to predict mortality rate over classes of risk) using goodness-of-fit as assessed by the Hosmer-Lemeshow statistic.

Results Data from 7703 patients were available for analysis. There was a decrease in overall hospital mortality, from approximately 19% at the beginning of the period of observation to approximately 12% at the end. There was also a decrease in the hospital standardised mortality ratio from 0.94 (95%CI 0.82–1.06) to 0.66 (95%CI 0.55–0.76). In general, both the APACHE II score and risk of death model performed adequately in each year with ROC curve AUCs of >0.8, albeit with progressively poorer performance over time and ‘model fade’ that approached statistical significance. There was progressively poorer calibration with the APACHE II risk of death model as indicated by the Hosmer-Lemeshow statistic, with a statistically significant difference between the predicted and observed mortality from 2003 onwards. Overall, there was moderately poor model performance in the diagnostic groups with the largest number of patients (sepsis and post-surgical complications).

Conclusion This study shows the progressively worse performance of the APACHE II illness severity scoring system over time due to ‘model fade’. This is especially so in common diagnostic categories, making this a clinically relevant finding. Future approaches to illness severity scoring should be tested and compared, such as re-estimating coefficients of the APACHE II diagnostic categories or using locally developed ones, moving to later evolutions of the system such as the APACHE III or APACHE IV, or developing novel artificial intelligence approaches.
Illness severity scoring systems such as the Acute Physiological and Chronic Health Evaluation (APACHE) have become important tools for the evaluation and planning of intensive care practice patterns. These systems objectively estimate patient risk for mortality from acute physiological and chronic health status. They are not, however, a tool used for deciding treatment for individual patients; they are a group measurement used for patients who have similar disease processes. Their origin in the late 1970s and early 1980s was driven by the need to relate such practice patterns to patient outcomes.

In the modern setting, tools such as the APACHE scoring system allow researchers and clinicians to quantify patient illness severity with a greater degree of accuracy and precision, which is essential for benchmarking and program evaluation. The interest in illness severity scoring systems is evidenced by the extensive body of literature that continues to advance both technical aspects of the systems themselves, and the applications for which they are used.

Middlemore Hospital was one the earlier facilities in New Zealand to implement APACHE II scoring in clinical settings. The routine scoring of patients began in the Intensive Care Unit (ICU) in 1986. This advance was facilitated in no small part by the availability of one of the developers of the APACHE system, Dr Jack Zimmerman, who spent an extended sabbatical in New Zealand, some of which was at Middlemore Hospital.

Despite enthusiastic support for APACHE II scoring by international opinion leaders at the time, the relevance and utility in New Zealand has been questioned from an early stage. The external validity of the system in such a different population from that in is developed was acknowledged by Zimmerman et al:¹

The NZ hospitals designated 1.7% of their total beds for intensive care compared to 5.6% in the US hospitals. The average age for NZ admissions was 42 compared to 55 in the US (p<0.0001). The NZ ICUs admitted fewer patients with severe chronic failing health (NZ 8.7%, US 18%) and following elective surgery (NZ 8%, US 40%). Approximately half the NZ admissions were for trauma, drug overdose, and asthma while these diagnoses accounted for 11% of US admissions. When controlled for differences in casemix and severity of illness, hospital mortality rates in NZ were comparable to the US. This study demonstrates substantial differences in patient selection among these US and NZ.

Furthermore, after more than two decades of use, it is unclear whether the performance of the APACHE II has been maintained. Patient casemix in New Zealand has changed from earlier times, and the improvements in supportive care that are now available may have decreased mortality for any given illness severity.

International opinion leaders are in general moving towards the more recently developed scoring systems such as APACHE versions III and IV, which have been shown to outperform older versions in studies of North American and European ICU populations.²

This paradigm is slowly translating to clinical practice in this part of the world: the Australian and New Zealand Intensive Care Society adult patient database now collect data sufficient to model both APACHE II and III scores.³
There are three aims of this study. We aim to:

- Assess change in APACHE II scores and hospital standardised mortality ratio at our ICU over a 9-year period from 1 January 1997 to 31 December 2005;
- Assess for changes in the performance of the APACHE II scoring system in predicting patient hospital mortality over the same period; and
- Assess for any clinical subgroups in which APACHE II scoring was particularly inaccurate or imprecise.

**Methods**

**Study population and setting**—Middlemore Hospital is the main hospital within the Counties Manukau District Health Board (CMDHB). The hospital serves a large urban population. The district catchment includes Manukau City which is rapidly expanding: the population has grown from 356,006 in 1996 to 454,655 at last census in 2006. The population can be summarily characterised as being young, multi-ethnic, and of low socioeconomic status compared with the rest of New Zealand. Middlemore Hospital is a tertiary referral centre for plastic surgery, burns, orthopaedics, and a range of medical sub-specialities. Any patient requiring neurosurgical or cardiothoracic surgical intervention is referred on to Auckland City Hospital as Middlemore Hospital does not have these facilities; all other patient categories remain at Middlemore Hospital.

Although there is a specialist regional paediatric hospital in the area, Middlemore Hospital is also a paediatric hospital; the Middlemore Hospital ICU therefore cares for those children down to 2 kg weight requiring intensive care accounting for approximately 120 paediatric admissions per year. The hospital is academically affiliated and thus a teaching institution. Middlemore Hospital has had between 700 and 900 acute beds over the time in which this research was done, and now also includes a satellite surgical centre which caters for the majority of elective cases apart from those that are particularly high risk.

Currently, the Middlemore ICU is nominally a seven funded-bed Level 3 facility. Since the inception of the Middlemore Hospital ICU in the late 1960s, the unit has been structurally modified on several occasions. As a result of the both national and local changes in healthcare strategy, the unit had at times had nominated HDU beds, and at other times not. Since 2004, there has been a four-funded bed Level I intensive care unit at a satellite surgical centre, which currently shares clinical governance, staff, policies and procedures with the main ICU at Middlemore Hospital. These patients were not included in this study.

**Data source**—All data were sourced from a single-centre relational database that has been in continuous use at the Middlemore Hospital ICU since January 1986. The database contains information on all patients admitted to ICU during this period, using data that is prospectively collected, collated, and agreed upon by senior specialists and the charge nurse at the time.

Data collection was progressively expanded during this period to ultimately include demographic information, APACHE II score, diagnostic information, ventilatory and inotropic support, procedures performed, and patient outcome. Patients who were less than 15 years of age, or who had been admitted solely for the purpose of a procedure such as difficult central venous line or endoscopy were not scored, as the system was not devised for these groups. The database specifically includes both patient death at both ICU and hospital discharge.

The database includes locally developed diagnostic codes (“adclasses” and “subclasses”) in addition to the APACHE II ones, which were developed to better reflect and discriminate disease categories related to the local population (see Appendix). Generic APACHE II diagnostic codes do necessarily provide a realistic reflection of the local disease categories and population outcomes. They can be ‘localised’ by adjustments to either disease categorisation and / or the category weights subsequently used with the APACHE II scores for calculating risk of death supported in the case of Middlemore Hospital by Zimmerman et al who emphasised differences between North American and New Zealand ICU patient populations.
Data were prospectively stored in Microsoft Access (Microsoft Corporation, Seattle, WA, USA), and retrospectively abstracted for analyses from a 9-year period from 1 January 1997 to 31 December 2005.

**Calculation of APACHE II scores and risk of death**—All APACHE II scores and risk of death were calculated at patient hospital discharge using the prospectively stored data and the logistic regression equation developed by Knaus et al. The data for calculation of the APACHE II score included physiological measurements in the first 24 hours of ICU admission, age and chronic health status. The APACHE II risk of death is calculated not only from scores but also diagnostic categories, which were rigorously and continuously evaluated by the senior ICU medical staff during the process of prospective data collection. Such minimisation of misclassification was necessary to avoid error arising from the heavy reliance of the APACHE II risk of death formula on reason for ICU admission.

**Statistics**—Standard statistics were used to describe data, making particular use of median and interquartile range to avoid assumptions around data distribution. Hypothesis testing was undertaken using Kruskal-Wallis equality-of-populations rank test for continuous variables, and the Pearson's Chi-squared test for categorical ones.

Risk-adjusted mortality by year was assessed by hospital standardised mortality ratios and 95% confidence intervals (regarding observed mortality as a binomial variable), which were obtained by dividing the number of observed hospital deaths in each year by the number of predicted ones using the APACHE II system.

Overall predictive performance of the APACHE II scoring system by year was gauged through discrimination (ability to discriminate between the patients who will die or survive at hospital discharge) and calibration (ability to predict mortality rate over classes of risk). Discrimination was assessed using receiver operating characteristic (ROC) curves, which plot the true positive rate (sensitivity, or predicted hospital deaths / observed hospital deaths) against the false positive rate (1-specificity, or 1-predicted hospital deaths / observed hospital deaths). The predictive performance is indicated in this method by the ROC area under the curve (AUC), with a value of 0.5 equating to random prediction and a value of 1.0 equating to perfect discrimination. The slope the curve indicates ratio of true positives and false positives, which also is known as the likelihood ratio. For the analyses in this article, equality of ROC AUC for each year of study was compared.

Calibration was assessed using the correspondence between the number of observed hospital deaths and the number of predicted hospital deaths within each 10% stratum (decile) of the cohort’s expected risk of death. The predictive performance is indicated in this method by goodness-of-fit as assessed by the Hosmer-Lemeshow statistic. The predictive performance of the APACHE II scoring system in major clinical subgroups was assessed by discrimination using hospital standardised mortality ratios within each of the major “adclasses”.

All analyses were performed using Microsoft Excel (Microsoft Corporation, Seattle, WA, USA) and Intercooled Stata 9.2 (Statacorp, College Station, TX, USA) software.

**Ethics**—The need for formal approval for the research process was waived by the National (New Zealand) Health and Disability Ethics Committee under the provisions made for clinical audit.

**Results**

Data from 7703 patients were available for analysis. Baseline patient characteristics are presented in Table 1. Numbers of patients admitted to the ICU increased steadily from 686 in 1997 to 730 in 2005. The demographic characteristics of patients changed over the period of observation, with a trend to older and more Māori patients.

There has also been a change in casemix of patients, with a reduction in the number of patients with diagnoses of poisoning and trauma, and an increase in the number of patients admitted after elective or emergency surgery. Patient length of stay has progressively reduced, as has the proportion of patients requiring mechanical ventilation. Overall hospital mortality decreased from approximately 19% at the beginning of the period of observation to approximately 12% at the end.
Table 1. Baseline patient characteristics by year

<table>
<thead>
<tr>
<th>Year</th>
<th>Age (mean)</th>
<th>Gender M/F</th>
<th>Ethnicity</th>
<th>Diagnostic groups</th>
<th>Median length of stay (hrs)</th>
<th>% mortality</th>
<th>% mechanical ventilation</th>
<th>% inotropic support</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>45.5</td>
<td>413/338</td>
<td>Local data not available</td>
<td>Surgical</td>
<td>15.3% 14.4% 14.1% 10.1% 45.8%</td>
<td>24.5</td>
<td>19%</td>
<td>65.24% 33.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Trauma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sepsis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Poisoning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>47.3</td>
<td>529/352</td>
<td>Eur 56% M 18% PI 19% Other 7%</td>
<td>Surgical</td>
<td>24.5% 17.0% 12.3% 5.5% 40.7%</td>
<td>22.0</td>
<td>17.8%</td>
<td>56% 23.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Trauma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sepsis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Poisoning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999</td>
<td>46.2</td>
<td>454/336</td>
<td>Eur 56% M 19% PI 15% Other 10%</td>
<td>Surgical</td>
<td>26.2% 15.7% 12.3% 7.1% 40.7%</td>
<td>23.0</td>
<td>13.0%</td>
<td>60% 22.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Trauma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sepsis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Poisoning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>49.3</td>
<td>469/390</td>
<td>Eur 53% M 21% PI 19% Other 7%</td>
<td>Surgical</td>
<td>31.6% 12.9% 9.9% 5.1% 40.5%</td>
<td>23.0</td>
<td>14.2%</td>
<td>57.3% 21.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Trauma</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sepsis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Poisoning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>45.1</td>
<td>487/328</td>
<td>Eur 48% M 20% PI 16% Other 16%</td>
<td>Surgical</td>
<td>29.7% 13.5% 16.0% 6.1% 4.6%</td>
<td>25.0</td>
<td>15%</td>
<td>54.8% 25.0%</td>
</tr>
<tr>
<td>Year</td>
<td>% Total</td>
<td>Injuries</td>
<td>Eur</td>
<td>M</td>
<td>PI</td>
<td>Other</td>
<td>Surgical</td>
<td>Trauma</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>----------</td>
<td>-----</td>
<td>---</td>
<td>----</td>
<td>-------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td>2002</td>
<td>48.2</td>
<td>487/328</td>
<td>Eur 48%</td>
<td>M 21%</td>
<td>PI 20%</td>
<td>Other 11%</td>
<td>Surgical Trauma Sepsis Poisoning Other</td>
<td>29.6%</td>
</tr>
<tr>
<td>2003</td>
<td>43.8</td>
<td>507/427</td>
<td>Eur 50%</td>
<td>M 22%</td>
<td>PI 21%</td>
<td>Other 7%</td>
<td>Surgical Trauma Sepsis Poisoning Other</td>
<td>27%</td>
</tr>
<tr>
<td>2004</td>
<td>52.7</td>
<td>424/351</td>
<td>Eur 51%</td>
<td>M 21%</td>
<td>PI 19%</td>
<td>Other 9%</td>
<td>Surgical Trauma Sepsis Poisoning Other</td>
<td>34.3%</td>
</tr>
<tr>
<td>2005</td>
<td>55.9</td>
<td>507/355</td>
<td>Eur 48%</td>
<td>M 25%</td>
<td>PI 19%</td>
<td>Other 8%</td>
<td>Surgical Trauma Sepsis Poisoning Other</td>
<td>32.8%</td>
</tr>
</tbody>
</table>

Eur=European, M=Māori, PI=Pacific Islander.
Figure 1. APACHE II scores and risk scores by year, presented as boxplots

Note: In these plots, the middle horizontal line represents the median; the box the second and third quartiles; and the whiskers the upper and lower extreme values which are no more than 1.5 × the interquartile range beyond the middle quartiles.

Figure 2. Hospital-standardised mortality ratio and 95% confidence intervals, by year
The APACHE II score decreased marginally over the period of observation as illustrated in Figure 1, with a median value of 14 in 1997 (IQR 9–21) and a corresponding value of 13 in 2005 (IQR 9–21). Although this reduction did achieve statistical significance (p=0.0001), it cannot be regarded as being clinically important. APACHE II predicted risk of death has remained stable over the period of observation, with a minor trend to reduction that did not achieve statistical significance (p=0.11).

The hospital-standardised mortality ratio decreased over the period observation as illustrated in Figure 2, with a value of 0.94 (95% confidence intervals 0.82–1.06) in 1997 and a corresponding value of 0.66 (95% confidence intervals 0.55–0.76) in 2005. Model adequacy for discrimination by APACHE II score is illustrated by year in Figures 3 and 4. In general, the APACHE II score performs adequately in each year with ROC curve AUCs of >0.8. However, there is deteriorating accuracy of mortality predictions over time (otherwise known as ‘model fade’\textsuperscript{10} that approaches statistical significance.

Corresponding model adequacy for discrimination by APACHE II predicted risk of death is illustrated in Figures 5 and 6. The risk model performs similarly to the APACHE II score showing a like degree of ‘model fade’.
Figure 3. ROC curves for APACHE II Score, by year

![ROC curves for APACHE II Score, by year](image)

**Note:** The predictive performance is indicated by the ROC area, with a value of 0.5 equating to random prediction and a value of 1.0 equating to perfect discrimination.

Figure 4. ROC curve AUC (95% CI) for APACHE II Score, by year, as shown in figure 3

![ROC curve AUC (95% CI) for APACHE II Score, by year](image)

**Note:** Marker labels indicate the P value for the test of equality of ROC areas relative to the reference year of 1997.
Figure 5. ROC curves for the APACHE II Risk score by year

Note: The predictive performance is indicated by the ROC area, with a value of 0.5 equating to random prediction, and a value of 1.0 equating to perfect discrimination.

Figure 6. ROC curve AUC (95% CI) for APACHE II Risk Score by year as shown in figure 5. Marker labels indicate the P value for the test of equality of ROC areas relative to the reference year of 1997.
Figure 7. Calibration curves for APACHE II predicted risk of death, by year showing the number of observed and predicted deaths within each 10% stratum (decile) of the cohort’s expected risk of death. Predictive performance is assessed by the Hosmer-Lemeshow statistic (see table 2)

Table 2. Model adequacy for calibration by APACHE II predicted risk of death, by year as indicated by the Hosmer-Lemeshow goodness-of-fit statistic for each of the calibration curves in figure 7. A high Hosmer-Lemeshow statistic and a P value <0.05 indicates poor correspondence between the number of observed and predicted deaths within each 10% stratum (decile) of the cohort’s expected risk of death

<table>
<thead>
<tr>
<th>Year</th>
<th>Hosmer-Lemeshow Goodness-of-fit Statistics (P values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>9.82 (0.132)</td>
</tr>
<tr>
<td>1998</td>
<td>11.14 (0.084)</td>
</tr>
<tr>
<td>1999</td>
<td>5.49 (0.482)</td>
</tr>
<tr>
<td>2000</td>
<td>3.31 (0.769)</td>
</tr>
<tr>
<td>2001</td>
<td>10.24 (0.111)</td>
</tr>
<tr>
<td>2002</td>
<td>8.04 (0.235)</td>
</tr>
<tr>
<td>2003</td>
<td>25.31 (0.0003)</td>
</tr>
<tr>
<td>2004</td>
<td>21.49 (0.001)</td>
</tr>
<tr>
<td>2005</td>
<td>19.41 (0.004)</td>
</tr>
</tbody>
</table>
Figure 8. Hospital-standardised mortality ratio (observed/predicted hospital deaths) for clinical diagnostic subgroups (“adclasses” as described in Appendix)

Model adequacy for calibration by APACHE II predicted risk of death is illustrated by year in Figure 7. There is progressively poorer goodness-of-fit as indicated by the Hosmer-Lemeshow statistic, with a statistically significant difference between the predicted and observed mortality from 2003 onwards as shown in Table 2.

Figure 8 illustrates model adequacy for discrimination by APACHE II predicted risk of death, according to clinical diagnostic subgroup. Although model adequacy was poorest in patients with neurological failure, there were only a small number of patients in this group. In contrast, the large number of patients with sepsis, respiratory failure, postoperative status, and circulatory failure makes the moderately poor model adequacy in these clinical subgroups more clinically relevant.

Discussion

Our data show that there has generally been a change in the overall casemix of patients admitted to the Middlemore Hospital ICU, with a decrease in the number of patients with poisonings and trauma over the period of observation, and an increase in those with complications as a result of surgery.

APACHE II scores have remained fairly constant over the period of observation, with only a subtle trend to decreasing patient illness severity that did not achieve statistical significance. The data also show that there has been with a reduction in crude and risk-adjusted mortality, as assessed mortality rates and by hospital standardised mortality ratios. Despite this, there has been a steady drop in the proportion of patients receiving mechanical ventilation over the period of observation, and the average length of patient stay.

Correlation between mechanical ventilation and increments in length of patient ICU stay has been noted in other studies. This change in outcomes and practice pattern
may reflect the benefits of clinical pathways within our hospital, and the earlier
detection and correction of physiological derangements that occurs in the modern,
more pro-active approach to provision intensive care.

An alternative, more pessimistic view is that this scenario may reflect earlier
discharges from our ICU to accommodate increasing demand in a setting of
increasingly limited resources. Reassuringly, if this latter scenario is the true one, then
outcomes appear to have been maintained despite this.

The data are in general terms consistent with a recent paper by Moran et al reporting
on intensive care outcomes using an international Australian and New Zealand ICU
database (ANZICS database), which to date has not included data from Middlemore Hospital and can therefore be regarded as independent. These investigators reported
an improvement in overall risk-adjusted mortality over the last 11 years, which they
did not attribute to any one specific factor.

Most medical administrators and practitioners would consider these improved
outcomes to be in some part causally related to corresponding improvements in
clinical care and therapeutic interventions. It would, however, take a more complex
minimum dataset than both the ANZICS database and our local one to study this
question appropriately.

There are two major findings of this study relating to the predictive performance of
the APACHE II system. The first is that there has been progressive deterioration
model adequacy in terms of both discrimination and calibration. Predictive
performance is generally acceptable when ROC curve AUCs are >0.8, and using these
and similar criteria it seems that continuing use this system in our current practice
may be unreasonable. The second is that the APACHE II system has been better
sustained in some clinical diagnostic subgroups but not others.

As is common to most ICUs, the largest clinical diagnostic subgroup in our dataset is
sepsis and post-surgical complications, and the APACHE II system has moderately
poor model adequacy in this subgroup, with prediction error of between 25-50%. Of
note, the subgroups with the largest prediction error in our dataset constitute only
~10% of the entire Middlemore ICU population.

The finding of ‘model fade’ over time is also consistent with those of Moran et al,
who demonstrated deteriorating model adequacy for the APACHE II system over
time, both in terms of both discrimination and calibration. This was the case even
after the authors recalibrated the APACHE II model by re-estimating coefficients for
the Australasian population, thereby optimising discrimination and calibration.

This is an important subtlety, since the performance of all illness severity scoring
models is well known to be poorer in populations that are different from those in
which they have been developed. This simple recalibration adjusts for geographical
differences in measured patient characteristics (physiology and diagnosis), although it
does not consider ICU characteristics and different organizational characteristics of
healthcare systems as a predictive variable. The Intensive Care National Audit and
Research Centre (ICNARC) model is in essence an adaptation of the APACHE model
that was developed by Rowan et al. in the 1990s in the United Kingdom, but over
the years has resulted in a completely independent model that is widely used in the
UK.
Opinion leaders now recommend regular recalibration of illness scoring systems to local and more contemporary cohorts,\textsuperscript{12} although to our knowledge there is no consensus or even propositions concerning thresholds for model performance that would trigger the recalibration process, or standardised methodology around the recalibration itself.

‘Model fade’ and poor model performance in diagnostic subgroups have led to the evolution of existing scores into a third and fourth generations of illness severity scoring systems, such as SAPS III and APACHE III and IV.\textsuperscript{2,12} The evolution of these scores did not involve simple recalibration of models by re-estimating coefficients, and instead involved the application of new statistical methods, the addition of new variables, an increase in the number of diagnostic groups, and a change to the measurement of certain physiological and diagnostic variables.

These scores can be expected to perform better as a result of their development in a cohort that is more contemporary and externally valid in terms of casemix, and also by using clinical information that was not initially taken into consideration during the development of the earlier systems.

There is a widespread move amongst ICUs to this newer generation of illness scoring systems, although their performance is only marginally better than earlier versions of the scores that have been more simply recalibrated by re-estimating coefficients.\textsuperscript{13} Notwithstanding, the APACHE III system is currently used more widely in the USA, with demonstrably greater discrimination and calibration than the original APACHE II system.\textsuperscript{2} It is too early to say at this time whether more recent evolutions of these systems such as the APACHE IV and SAPS III systems will demonstrate continued improvement.

The findings of our study do not address one of the conundrums of illness severity scoring: the interpretation of changes in scores and outcomes over time. As with other studies, it is impossible to tell from our data whether our results are due to improved patient care and access to care, or alternatively from the deteriorating performance of scoring systems because of changing patient casemix.

Our cumulative clinical experience is in keeping with others: ICU patients are in general sicker than previously, with improving outcomes despite this. Confirmation of this perception will only be forthcoming with studies that extend data collection to include other indicators of patient illness severity and practice patterns, and the use of statistical approaches that use causal or structural times series modelling.

The strength of this study is its size and completeness. This study, running from 1997 to 2005 inclusive contains a large dataset over a nine year period without gaps. The major weaknesses of this study are those that are inherent to any scoring system that is dependent on clinical classification of patients into diagnostic categories (whether local diagnostic codes (‘adclasses” and “subclasses”) or APACHE II ones). There are no explicit criteria to improve consistency within or between ICUs in making these classification, and all due care was taken in our database to limit subjectivity and optimize accuracy and precision as much as possible.

In terms of the future of illness severity scoring, good reasons abound for us to persist with the APACHE scoring system at Middlemore Hospital, as opposed to moving to
others such as organ failure scoring systems (Multiple Organ Dysfunction Score, Sepsis-related Organ Failure Assessment).

The choice of method within any particular ICU is critically dependent on the degree of confidence in its use; the APACHE scoring systems are more validated than the other choices at Middlemore Hospital ICU. Moreover, it is our opinion that the APACHE scoring systems are also subject to rigorous remodelling and adaptation: this is essential to ensure that the system reflects changes in underlying characteristics of patients and healthcare delivery systems, and therefore correctly model the relationships with patients’ outcomes.\textsuperscript{2,14}

Notwithstanding, there have been encouraging results with loosely-termed ‘artificial intelligence’ approaches. Frize and Walker reported early success of their pilot of neural networks in both adult and neonatal intensive care.\textsuperscript{15} Investigation into these modelling methods may prove fruitful for the future, and may result in better performance although this is yet to be definitely demonstrated.\textsuperscript{16–18}

Our data indicate that we should be preparing to move forward from the APACHE II system. Three workstreams are suggested by the results of this study, which should probably be run concurrently with the results determining the final solution for illness severity scoring.

The first workstream should involve recalibration of the APACHE II model by re-estimating coefficients for our local population using local diagnostic codes (“adclasses” and “subclasses”) and/or APACHE II one s. The second should involve a trial of the APACHE III system. The third should involve a pilot of artificial intelligence approaches.

The performance of these three approaches in our population should determine which illness severity scoring system should be used in short and medium term. However, it would appear that regular re-calibration should be undertaken irrespective of what model is chosen, in order to minimise ‘model fade’ and provide clinicians and managers interested in benchmarking a well validated model to predict mortality.

**Competing interests:** None known.

**Author information:** Susan L Mann, Department of Intensive Care Medicine, Counties Manukau District Health Board, Manukau, South Auckland; Mark R Marshall, Nephrologist, Department of Internal Medicine, Counties Manukau District Health Board, Manukau, Auckland; Alec Holt, Director Health Informatics Programme, Department of Information Science, University of Otago, Dunedin; Brendon Woodford, Department of Information Science, University of Otago, Dunedin; Anthony B Williams, Intensivist, Department of Intensive Care Medicine, Counties Manukau District Health Board, Manukau, South Auckland

**Acknowledgements:** The authors thank Mr Mpatisi Moyo (Decision Support, Middlemore Hospital) and Mr Gary Jackson (Public Health Physician, Counties Manukau District Health Board).

**Correspondence:** Susan L Mann, PO Box 25-075, St Heliers, Auckland 1740, New Zealand. Fax: +64 (0)9 2760034; email: smann@xtra.co.nz
References:


Appendix on next page
Appendix ("Adclass and subclass" classification)

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRAuma (admitted to ICU for trauma or direct complication of trauma)</td>
</tr>
<tr>
<td>2</td>
<td>POIsoning (admitted to ICU because of poisoning)</td>
</tr>
<tr>
<td>3</td>
<td>AIRway disease</td>
</tr>
<tr>
<td>4</td>
<td>ANAphylaxis</td>
</tr>
<tr>
<td>5</td>
<td>ASPhyxiation</td>
</tr>
<tr>
<td>6</td>
<td>ASThma</td>
</tr>
<tr>
<td>7</td>
<td>SEPsis (clinical sepsis plus defined focus or positive blood culture)</td>
</tr>
<tr>
<td>8</td>
<td>OBStetric complication</td>
</tr>
<tr>
<td>9</td>
<td>CPR (post)</td>
</tr>
<tr>
<td>10</td>
<td>SURgery</td>
</tr>
<tr>
<td>11</td>
<td>CIRculatory failure</td>
</tr>
<tr>
<td>12</td>
<td>CNS failure</td>
</tr>
<tr>
<td>13</td>
<td>GIFailure</td>
</tr>
<tr>
<td>14</td>
<td>METabolic failure</td>
</tr>
<tr>
<td>15</td>
<td>NEUromuscular failure</td>
</tr>
<tr>
<td>16</td>
<td>RENal failure</td>
</tr>
<tr>
<td>17</td>
<td>RESpiratory failure</td>
</tr>
<tr>
<td>18</td>
<td>PROCedures admission</td>
</tr>
<tr>
<td>19</td>
<td>MIScellaneous</td>
</tr>
</tbody>
</table>

Subclass on next page
1. TRA type of trauma
   - BLUnut
   - PENetrating
   - BURN

2. POI type of poisoning
   - CYClic medication (+sedative)
   - SEDative medication
   - MEDication (other)
   - NON-medicinal

3. ANA type of anaphylaxis
   - MEDicinal
   - NON-medicinal

4. ASP type of asphyxiation
   - DROWning
   - HANGing
   - STRangulation
   - NON-medicinal

5. SEP locus of sepsis
   - BLOod only
   - ENDOCardium
   - GENital tract
   - GIT tract
   - JOInt
   - MENinges
   - RESp tract
   - SOFT tissues

6. GIF type of GI failure

7. MET type of metabolic failure

8. NEUromuscular failure

9. SUR type of surgery

10. REN type of renal failure

11. CIR type of circulatory failure
   - DYSrhythmia
   - CCU netrating
   - PULmonary embolism
   - AMI (acute)

12. CNS type of CNS Failure

13. GIFT type of GI failure
   - HAEmorrhage
   - HEPatic failure
   - PANcreatitis

14. MET type of metabolic failure
   - HEAt stroke
   - HYPothermia
   - DIAbetic

15. NEUromuscular failure
   - MYAsthenia
   - MISCellaneous
   - GBS (Guillain-Barré)

16. REN type of renal failure
   - ARF (acute)
   - CRF (chronic)
### 17. RES type of respiratory failure 18 PROCedure type admitted for

<table>
<thead>
<tr>
<th>NECK</th>
<th>CVP insertion</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORThopaedic</td>
<td>Dialysis</td>
</tr>
<tr>
<td>PLAstic</td>
<td>OTHer</td>
</tr>
</tbody>
</table>

OSA

<table>
<thead>
<tr>
<th>CORD</th>
</tr>
</thead>
</table>

MISCELLANEOUS