Scoring and outcome audit systems relevant to emergency medicine

M. WATERS & P. NIGHTINGALE

Department of Accident and Emergency Medicine, University Hospital of South Manchester, Manchester

Scoring systems in medical practice are used to measure deficits of physiology and function in order to provide standard reference points immediately recognizable by other observers, and from which response to treatment can be measured.

In 1974, Teasdale & Jennett introduced the Glasgow Coma Score (GCS), which is now internationally accepted as a means of describing impaired conscious level. The three component sections of eye opening, motor and verbal response have maximum scores of 4, 6 and 5 respectively, giving a best score of 15. The verbal component has been modified for use in very young children (ATLS Manual, 1989) as follows: normal words, smiles or follows with eyes 5; cries but consolable 4; persistently irritable 3; restless and agitated 2; no response 1. For intubated patients, the Intensive Care Society recommend that the GCS should be calculated but a note is made of whether the patient is sedated and/or paralysed. The verbal component has been adjusted to account for this: appears orientated and able to converse 5; responsive but ability to converse is in doubt 3; generally unresponsive 1. The GCS is used in its own right and as an integral part of other scores such as Revised Trauma Score (RTS), Simplified Acute Physiology Score (SAPS) and Acute Physiology And Chronic Health Evaluation (APACHE).

The need to audit existing levels of pre-hospital and trauma care has become a necessity, and not an option. In the USA, where scoring systems were first used widely, it is an absolute requirement of a regional trauma centre that there should be an established system for outcome audit. Some simple observation scores have been used by paramedics to decide on the most appropriate treatment and level of care or institution that each patient requires. One of the first of these was the Trauma Index (Kirkpatrick et al., 1971) which has been largely replaced by the CRAMS scale and Pre-hospital Index. The CRAMS scale is a basic assessment of the effects of injury on the Circulation, Respiration, Abdomen, Movement and Speech in terms of severe abnormality through to normality (Gormican, 1982). These score 0–2 respectively, and a maximum score of 10 indicates normality. The Pre-hospital Index is the sum of coded values (0–5) for defined ranges in heart rate, and systolic blood pressure, together with
respiratory effort and conscious level (Koehler et al., 1986). An extra 4 points are given for the presence of penetrating injury to the abdomen or chest. The total score range is 0–24, with a score of 0 indicating normality and 4 taken as the decision level.

The Trauma Score (Champion et al., 1981, 1986) has gained a wider acceptance for several reasons. It is derived from a coded value for the respiratory rate and effort, capillary refill time, systolic blood pressure and GCS. The maximum score (16) indicates normality. Subsequent experience with the Trauma Score led to its refinement with omission of respiratory effort and capillary refill weightings to give the Revised Trauma Score with a maximum of 12 points (Boyd et al., 1987). Both scores are inversely related to the Injury Severity Score (ISS) and a probability of survival can be made from each of them.

The Abbreviated Injury Scaling dictionary (AIS) is a comprehensive list of injuries which have been scored from 1 to 6 (6 being defined as 'virtually unsurvivable') by the American Association for the Advancement of Automotive Medicine. To derive the Injury Severity Score (ISS), the body is divided into six anatomical regions and all of the injuries sustained are noted at the time of death or discharge from hospital. The score is then the sum of the square of the highest AIS score in each of the three worst affected body regions. Any injury with an AIS of 6 automatically converts to a maximum ISS of 75. Analysis of the outcome of many thousands of patients and their age, in relation to the injuries sustained has given the Trauma score/Injury severity score (TRISS) methodology (Boyd et al., 1987) which developed into the major trauma outcome study (MTOS). The RTS can be plotted against ISS, and LD50 lines produced for patients over and under 55 years, and for blunt and more recently penetrating injuries as well. MTOS has now spread internationally. To facilitate comparisons with other units, a Z statistic can be produced, which gives an indication of overall performance ranked against US trauma centres, and an M statistic, which is an indication of the local population match to the baseline population. Both are required for a correct assessment of performance.

In an effort to predict post-operative complications following penetrating abdominal trauma, Moore and colleagues (Moore et al., 1981) utilized a simple modification of the AIS with an estimate of degree of intra-abdominal injury found at laparotomy. Values for the Penetrating Abdominal Trauma Index (PATI) of greater than 25 were associated with a marked increase in post operative complications.

Other scoring systems have been developed which are more appropriate to the Intensive Care setting. As the treatment in Intensive Care Units developed, and became more expensive, a need arose to audit the number and type of therapeutic interventions undertaken for each patient. The therapeutic intervention scoring system (TISS) was introduced in 1974 by Cullen and updated in 1983 (Cullen et al., 1974; Keene et al., 1983). Knaus has described 71 interventions each scoring 1 to 4 points as markers of basic nursing care, monitoring which is technology intensive, monitoring which is personnel intensive, and active treatments respectively. In the USA, information derived from daily scores performed on each patient can relate the degree of therapy when compared to outcome, can give an assessment of the number and experience of nursing staff required to care for a specific type of patient, describes length of stay patterns and give an estimate of the best time for discharge to the ward (when the TISS is <10).

The Acute Physiology and Chronic Health Evaluation (APACHE) scoring system
has been developed in Washington DC (Knaus et al., 1981). The aim was to develop a physiologically-based system that would enable patients to be stratified by severity of illness, thus eliminating susceptibility bias when comparing small groups of patients, since randomization might be difficult to achieve and ethically questionable in critical illness. It was envisaged that multi-centre and international comparisons of data would be possible. The score itself is the sum of three components: the acute physiology score (APS), a coded value for age and previous chronic ill health. The nature of the admission (medical, surgical, elective or emergency) is also taken into account. The APS was derived from 34 variables and each was assigned 0–4 points for the degree of deviation from the normal range.

The APS is the sum of these for the worst value of each variable in the first 32 h of ICU admission. The higher the score, the sicker the patient is. There were four categories describing patients health through the preceding 6 months, ranging from ‘no functional impairment’ (class A) to ‘severe restriction of lifestyle by disease’ (class D). It was recognized that not all 34 variables were always measured and there was a possibility of undermeasurement. The score was validated against hospital outcome and the amount of therapy received. The probability of an in-hospital death increased with increasing scores. Of the pre-admission health status categories, only class D was found to be helpful in predicting short-term mortality, and then only for non-operative admissions.

Recognition of APACHE I as an unwieldy system led to the development of the Simplified Acute Physiology Score (SAPS) by Le Gall (Le Gall et al., 1983) and the APACHE II system (Knaus et al., 1985a). These reduced the number of measurements to be made to a mandatory 13 and 12, respectively. Both take age into account. SAPS was validated in 8 French ICUs, using the worst value of each in the first 24 h of ICU admission. A fixed value of 3 points was assigned to ventilated patients, rather than using direct measurements of oxygenation as in APACHE. A good correlation was found between SAPS and ICU (rather than hospital) mortality.

APACHE II also uses the worst in 24 h values for the APS, plus weightings for age and severe chronic organ system insufficiency in association with the type of admission. Non-operative (not admitted directly from theatre) and emergency surgical admissions with chronic ill health had an increased risk of death compared to elective surgical admissions, and were weighted more heavily. The maximum score is 71, but no patient in the validation group exceeded 55. In addition, each patient was assigned to one of 40 diagnostic categories where possible, or to one of five mutually exclusive organ system failure/insufficiency categories, i.e., cardiovascular, respiratory, gastrointestinal, neurological, or renal/metabolic. Each 5 point increase in APACHE score was associated with an increase in mortality rate, but the risk of an in-hospital death was also strongly related to the disease process. Coefficients were derived for each diagnostic category by multiple regression analysis. Diagnoses with very negative coefficients carry a better prognosis than those with very positive ones. This, and the APACHE score, are put into an equation which gives a risk of death (ROD) value for that patient after 24 h on ICU i.e., \( \log_{1-\text{ROD}} \text{ROD} = A + (B \times \text{APACHE II SCORE}) + \text{disease category coefficient} + \text{emergency surgery weighting} \), where A and B are constants. The results were analysed by a classification matrix, using a 0.5 risk as the decision criterion, and taking correct prediction of death as the positive result.
Typical classification matrix. Data from our Institution.

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>alive</td>
</tr>
<tr>
<td>Actual</td>
<td></td>
</tr>
<tr>
<td>alive</td>
<td>175</td>
</tr>
<tr>
<td>dead</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>247</td>
</tr>
</tbody>
</table>

Sensitivity = 36.8%  
Specificity = 95.6%

Predictive value positive = 84%  
Predictive value negative = 70.9%  
Overall correct prediction rate = 73.1%

This type of analysis is suitable for comparison of severity of disease with outcome, and also within disease categories. The latter method of comparing two groups is superior to the use of mean values to indicate similarity (Knaus et al., 1984, 1985a,b). It was stressed that the individual RODs should never be used on an individual basis to influence subsequent management. It was also felt that severity of illness might better be reflected by earlier data collection, ie., at the time of admission. A study on a subset of patients demonstrated that in 88% of physiological measurements, the worst in 24 h values were the same as the admission ones. Also, the use of admission values changed the APS by less than 5 points in 81% patients. A commitment to earlier data collection in the future was suggested (Knaus et al., 1985a).

Some patients were deliberately excluded from the validation studies for SAPS and APACHE II, namely burns patients, children, straightforward myocardial infarction and post coronary artery bypass patients. The authors advise caution in trying to apply APACHE to these groups and to patients with a specific diagnosis not listed among the given diagnostic categories. Fedullo found a poor correlation with outcome when the APACHE II system was applied to acute left ventricular failure patients (Fedullo et al., 1988).

Although APACHE successfully predicted the majority of patients to live or die, there was still a large number of false negatives (predicted to live but died) and a significant but smaller number of false positives (predicted to die but lived). It rapidly became apparent that in order to improve predictive accuracy, the APS should be repeated over a period of time, preferably accounting for the rate of change from day to day (Jacobs et al., 1987). Bion used the sickness score (SS) in this way. It is a modification of APACHE II devised to audit the problems encountered in transporting sick patients (Bion et al., 1985). They found that survival from severe illness was associated with a decrease in mean SS of 45% by day 4, whereas non-survivors improved by only 18% (Bion et al, 1988). A major problem was that the diagnostic category used on admission might not be the most appropriate by day 4, eg. from trauma to sepsis or ARDS.

Several studies from Saudi Arabia have concentrated on improving the accuracy of prediction of death with a view to identifying those patients who will not benefit from...
prolonged intensive care, being mindful of the vast increases in the costs of monitoring and treating patients on ICUs (Chang et al., 1986, Jacobs et al., 1988). Firstly, no patients on their unit died with a ROD < 7% and no patient survived with a ROD > 60% or an APACHE score > 35 points (Chang et al., 1988b). They postulated that patients whose RODs fell outside these limits would not benefit from remaining on an ICU after the initial 24 h of assessment. Secondly, if a system could be developed which when tested prospectively, could predict death with 100% accuracy relatively early on, then great savings could be made in terms of cost and of reduced stress on relatives if treatment were withdrawn earlier, as well as enabling scarce resources to be focused on other patients who might benefit from intensive care. The real danger of inducing self-fulfilling prophecies was noted. One way of achieving this is to raise the decision level above which all patients are predicted to die. The resultant rise in specificity is accompanied naturally by a fall in sensitivity, due to the increase in numbers of false negatives. In addition, the authors modified the APS considerably by using the best GCS values in each 24 h study period. The GCS is heavily weighted in the APS (coded value = 15 - GCS), so a reduction in the APS tends to help prevent the overprediction of death. They developed two systems using trend analysis of daily APACHE II scores, the second also included an organ failure score (Chang et al., 1988b,c). Both achieved 100% specificity. They were able to observe and calculate both absolute levels and patterns of response to treatment that were associated with death or survival, for the majority of patients. Outcome predictions using organ failure scores were the most accurate, (specificity 100%, sensitivity 51·4%, and total proportion correct 83·3%). They have recently estimated that there is a 1·4% risk of false prediction of death at 95% confidence level using this technique, and despite the sophistication of the above, still feel that there is a piece for the single day 1 APACHE score for general audit purposes.

Interestingly, Kruse et al., (1988) found no significant difference in accuracy of prediction of outcome between APACHE II scores and clinical assessment. The latter was best performed by second year residents and critical care nurses.

There have been two major international comparisons of intensive care, between France/USA (Knaus et al., 1982) and New Zealand/USA (Zimmerman et al., 1988), using APACHE II and TISS. They both showed important differences in the number of designated ICU beds, patient selection, length of stay, amount of therapeutic intervention and levels at which treatment is withheld. The outcome was similar in comparable institutions for cardiovascular, respiratory, and neurological organ system failures, but there was a lower mortality for gastrointestinal conditions in USA compared to France. Outcome was comparable between New Zealand and USA.

In contrast to the above methodology, extensive work by Shoemaker has shown that two important parameters which influence outcome in critically ill surgical patients are oxygen delivery (DO₂) and oxygen consumption (VO₂). Resuscitation to normal physiological values (ie. to a very low APACHE score) was not significantly associated with survival. When supra-normal oxygen deliveries were supplied, by means of controlled plasma volume expansion with blood or colloid plus inotropic support where necessary, there was a significant fall in morbidity and mortality, especially from acute renal failure. The observation that oxygen consumption rose with delivery appeared to reflect an oxygen debt as well as an increase in metabolic rate.
Van Lanschot et al., (1988) attempted to combine data on oxygen transport variables with SAPS and reported that there was no improvement in predictive accuracy, although his conclusions have been criticized (Shoemaker, 1989).

The Mortality Prediction Model (MPM) was designed for use on admission to ICU (Lemeshow et al., 1985). It has subsequently been modified to allow scores at 24 and 48 h. The parameters included in the score had each been identified as being significantly related to outcome by regression analysis, and are not all purely representative of physiological dysfunction, e.g. has CPR been performed prior to admission, is cancer part of the present problem, and is infection likely to be present. A mortality probability is derived from the score. It was intended as a tool for audit, rather than for patient selection, and when compared to SAPS and APACHE II, gave comparable results (Lemeshow et al., 1987).

Despite the technological advances that give more and better quality information per patient, Knaus found that the probability of an in-hospital death was significantly reduced where there was a full time director of ICU, good interaction and co-ordination of ICU staff, and good nursing continuity of care. Outcome in the 13 hospitals studied was not related to the aggressive use of invasive monitoring, amount of therapy used, or teaching hospital status! (Knaus et al., 1986).

CONCLUSION

Scoring systems should be brief and easy to do, since the major problem with data collection is always the human factor. The need to conduct medical audit in order to maintain and improve our clinical practice is essential. Scoring systems quantify and collate relevant data in a compact form that can be used to guide rather than dictate clinical management. However, individual predictions, whether by score or clinical assessment may give a result which has dire implications if followed blindly. Audit of group death rates by severity of disease and diagnosis should be of great benefit in improving standards of care, assessing the efficacy of new forms of treatment, as well as identifying anomalous deaths and survivors.

REFERENCES

Scoring and outcome audit systems


