Intensive Care Med (2005) 31:1618–1623
DOI 10.1007/s00134-005-2825-8

Jean-Roger Le Gall

The use of severity scores in the intensive care unit

Received: 13 September 2005
Accepted: 13 September 2005
Published online: 22 October 2005
© Springer-Verlag 2005

J.-R. Le Gall (✉)
Department of Intensive Care Medicine,
Saint-Louis University Hospital,
Paris, France
e-mail: jr.legall@sls.ap-hop-paris.fr
Tel.: +33-1-42499421
Fax: +33-1-42499426

Background

Around 1980 several intensivists decided to score the severity of ICU patients in order to compare the populations and evaluate the results. The outcome of intensive care depends on several factors present on the first day in the ICU and on the patient's course under ICU therapy. The severity scores comprise usually two parts: the score itself and a probability model. The score itself is a number (the highest number, the highest severity). The probability model is an equation giving the probability of hospital death of the patients. This seminal comprise two parts: the classification of the scores and their practical use.

Classification of the severity scores

Many severity scores have been published but only a few are used. Most scores are calculated from data collected on the first ICU day; these include the Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), and Mortality Prediction Model (MPM). Others are repetitive and collect data every day throughout the ICU stay or for the first 3 days; these include the Organ System Failure (OSF), Organ Dysfunction and Infection System (ODIN), Sequential Organ Failure Assessment (SOFA), Multiple Organs Dysfunction Score (MODS), Logistic Organ Dysfunction (LOD) model, and Three-Day Recalibrating ICU Outcomes (TRIOS).

First-day ICU severity scores

Subjective scores

These scores are established by a panel of experts who choose the variables and assign a weight to each variable based on their personal opinion. For each variable a range of normality is defined, with a score of 0 within this range. The more abnormal the result, the higher the weight that is given, from 0 to 4 points. The total number of points constitutes the score. The most commonly used scoring system is APACHE II [1]. This includes variables such as age, preexisting diseases, and 12 acute physiological variables. This yields a probability of hospital death depending on the main diagnosis.

Objective scores

Development of a multipurpose probability model requires that a large database be compiled using data from many ICUs. Variables collected can generally be classified into four groups: age, comorbidities, physiological abnormalities, and acute diagnoses. Some systems have introduced variables designed to decrease the lead-time bias. The principal outcome for each of the systems is vital status at hospital discharge. Other outcome measures (e.g., vital status 28 days after hospital discharge or quality, life among long-term survivors) can also be modeled. Logistic regression modeling techniques, smoothing methods, and clinical judgment are used to select variables, determine ranges, and assign weights. All of the systems result in a logistic regression model that estimates the risk of death. In chronological order of publication the main objective scores are APACHE III [2], the SAPS II [3], and the MPM II [4].

APACHE III. This score uses largely the same variables as APACHE II but a different way in which to collect the neurological data, no longer using the Glasgow Coma Score. It adds particularly two important variables: the patient's origin and the lead-time bias. The acute diagnosis is taken into account; one diagnosis must be preferred.

SAPS II and the expanded SAPS II. The same technique was used to construct SAPS II. The database, however, was established from European and North American ICUs, and the acute diagnosis were not included. The authors considered it too difficult to select a single diagnosis for an ICU patient. As for other scoring systems the discrimination and particularly the calibration of the SAPS II model does not fit when applied to a new population. The model can be adapted to a country or a specific population by a customization process or by expansion of the model through the addition of new variables. For example, a revision of SAPS II has been proposed by Aergerter et al. [5]. Retrospective analysis of 33,471 prospectively collected multicenter data was performed in 32 ICUs located in the Paris aera. They developed two logistic regression models. The second one reevaluated items of SAPS II and integration of the preadmission location and chronic comorbidity. Another proposal was recently made by Le Gall et al. [6]. From a database of 77,490 admissions in 106 French ICUs they added six admission variables to SAPS II: age, sex, length of the ICU hospital stay, patient location before ICU, clinical category, and whether drug overdose was present. The statistical qualities of the expanded SAPS II are much better than those of the original and even the customized SAPS II. The original SAPS II mortality prediction model is outdated and needs to be adapted to current ICU populations. The original SAPS II may be used to score the ICU patients' severity. But to calculate the standardized mortality ratio or the ICU performance measure it is now necessary to use the expanded SAPS II Adding simple data, routinely collected, to the original SAPS II led to better calibration, discrimination, and uniformity-of-fit of the model. The statistical qualities of the expanded SAPS II are much better than those of the original and the customized SAPS II. Above all, the expanded SAPS II is easy to obtain from the existing databases. It is now the simplest system for precisely measuring ICU performance and comparing performance over years.

MPM II. In the case of the MPM II one has not a score but a model giving directly the probability of hospital death. This uses chronic health status, acute diagnosis, a few physiological variables, and some other variables including mechanical ventilation. The database is the same as that for the SAPS II. Four models have been proposed: MPM II at admission and at 24, 48, and 78 h.

SAPS 3. A worldwide database of 19,577 patients was used to develop SAPS III. It comprises three parts: chronic variables, acute variables including the sepsis and its characteristics, and physiology. The calculated probability of ICU and hospital death emerges by adding diagnoses to the model. Evaluation of ICU performance is adapted to each ICU according to its case-mix [7, 8].

Repetitive scores

Subjective scores

OSF. Data on five organ failures are included in the OSF system [9]. The main prognostic factors are the number and duration of these failures. Mortality is close to 100% when three organs failures persist for 5 days or longer.

ODIN. Fagon et al. [10] proposed the ODIN system in 1993. This includes data on six organ failures plus one infection and differentiates prognosis according to the type of failures.

SOFA. Published in 1998 by Vincent et al. [11], the SOFA subjective score was evaluated on 1,449 patients. Data on six failures are scored on a scale of 0–4. One failure plus a respiratory failure indicate the lowest mortality; all the other combinations yield a mortality between 65% and 74%. Subsequent analyses have considered the maximal score plus the maximal change and have shown that the latter has a lower prognostic value than the former.

Objective scores

MODS. In 1995 Marshall et al. [12] examined the definitional criteria of organ failures proposed in the literature and tested these criteria in a population of 692 patients. The result of their work, the MODS, comprises a score based on six failures each scored from 0 to 4. This considers the time of occurrence of each failure; respiratory failure was found to be the first (1.8±4.7 days) and hepatic failure the last (4.7±5.5 days). They showed that mortality depends non only on the admission score but also on course.

LOD model. This model based on the LOD is the only one based on logistic regression. From a European North American database 12 variables were tested and 6 organ failures defined [13]. The originality of the model is to give to each dysfunction a weight of 0–5 points. Severe neurological, cardiovascular, and renal failures are scored

5, severe respiratory failure 3, and severe hepatic failure 1. The model has been tested over time. The difference between the LOD scores on day 3 and day 1 is highly predictive of the hospital outcome.

TRIOS. A composite score using daily SAPS II and LOD score for predicting hospital hospitality in ICU patients hospitalized for more 72 h was proposed by Timsit et al. [14] in 2001. This TRIOS composite score has excellent statistical qualities and may be used for research purposes.

Model validation

Model performance must be demonstrated in a sample of patients independent of that used to develop the models. Validation samples have been assembled either by collecting data on a new cohort of patients or by randomly splitting an available database into two portions—one used to develop the model and the other to validate it [15].

Model calibration

Calibration evaluates the degree of correspondence between the estimated probabilities of mortality produced by a model and the actual mortality experienced by patients. Calibration can be statistically evaluated using formal goodness-of-fit tests [16]. What information does the assessment of calibration provide? If a model estimates that a set of patients have a probability of hospital mortality of 0.38, this means that among 100 such patients 38 would be expected to die and 62 to live. When the observed number of deaths is close to the number predicted by the model, it is considered to be well calibrated.

To test calibration formally patients are rank-ordered according to their probability of mortality and grouped into range-defined strata. Typically ten such strata are formed, each containing approximately the same number of patients (called "risk deciles"). To obtain the predicted number of deaths in a stratum, the probabilities of mortality for all patients in that stratum are summed. Formal goodness-of-fit testing compares the observed with the predicted number of deaths and the observed with the predicted number of survivors in each stratum of patients. The resulting value can be used to determine whether the combined discrepancy between observed and predicted outcome across all strata is within sampling variability. If differences are large, the model does not correctly reflect the outcome in that cohort of patients.

Model discrimination

Discrimination uses the area under the receiver operating characteristic (ROC) curve to evaluate the ability of a model to distinguish patients who die from patients who live, based on the estimated probabilities of mortality. To construct the ROC curve [17] a sequence of probability cutoff points is specified, and a 2×2 classification table of predicted and observed outcome is constructed for each cutoff. For example, if the cutoff is set at 0.35, any patient whose probability of mortality is 0.35 or higher would be predicted to die, whereas any patient whose probability is less than 0.35 would be predicted to live. Observed mortality is noted for each patient and from the resulting 2×2 table the false-positive and true-positive rates are determined. All these pairs of rates for the sequence of cutoff points are then plotted, resulting in the visual presentation of the ROC curve. The higher the true-positive rate is relative to the false-positive rate, the greater is the area under the ROC curve.

Interpretation of the area under the ROC curve is quite simple. If the entire sample were divided into patients who lived and patients who died, and each patient who lived were paired with each patient who died, there would be $n_1 \times n_0$ such pairs (where n_1 is the number of patients who lived and n_0 is the number who died). The area under the ROC curve is the proportion of the total number of pairs in which the model resulted in a higher probability for the patient who died than the patient who lived. Clearly, if the value is in the neighborhood of 0.50, the model performs no better than the flip of a coin. Developers of models are usually not satisfied unless the ROC area of a model exceeds 0.70.

Comparison of the models

Comparison provided by the developers

The latest generation of models (APACHE III, SAPS II, MPM II) have been evaluated by the developers. Ideally information would be available on calibration and discrimination in both the developmental and the validation samples. Except for the physiology component of APACHE III the system was developed using the entire sample, and therefore no independent validation sample results are reported in the publication which presents the system. Reported discrimination power of all three systems was excellent. In the total sample the area under the ROC curve was 0.90for APACHE III, 0.88 for SAPS II, and 0.84, 0.84, 0.81, and 0.79 for $MPM_{0,}$ MPM_{24}, MPM_{48}, and MPM_{72}, respectively, in the developmental samples. For SAPS II the area under the ROC curve was 0.86 in the validation sample and 0.82, 0.84, 0.80, and 0.75 in the validation samples for the four models of the MPM II. Information for evaluating the goodness-of-fit of APACHE III has been not reported. The calibration of the models in the SAPS II and MPM II systems indicated that all of the models fit the data well, as reflected by the close correspondence between the observed and predicted outcomes across the entire range of probabilities. Calibration was excellent in the developmental samples for all of the SAPS II and MPM II models, and close correspondence between observed and predicted numbers of deaths was noted in the independent validation sample as well.

The qualities of models over time

The case-mix does not remain the same as the therapies evolve over time, and the selection of patients admitted to ICUs may differ over time, and therefore published scoring systems become obsolete. Usually the ROC curves remain good, but the validation, when the scores are applied to other populations, is poor. Depending on the score it may be useful to customize it to the respective population. To compare patient groups in a clinical study it is not necessary to charge the score used. For instance the SAPS II continues to be used in many scientific publications. To evaluate the performance of an ICU it is better to customize the score. There are two ways in which to do this: change the probability equation or the weight of each variable [18], or add new variables, which requires a further collection of data.

Practical use of the scores

Scoring systems have been proposed in use for individual patient prediction to evaluate the performance of ICUs and for therapeutic trials. In general, proposed uses for scores or probabilities can be considered at both the individual patient level and the aggregate level. That is, one may use a score to make a statement about groups of patients. Serious consequences may arise depending on the action that one takes in response to such a statement, and therefore a conservative approach to the application of scores to individuals is necessary. After all the careful research that has produced the various severity scoring systems, the uses to which they can be appropriately be put are still not universally agreed [19]

Prediction for individuals patients

The systems can be used either to determine objective risk of death or in a clinical assessment. Meyer et al. [20] showed that among the patients who were predicted by both methods to die, more than 40% of actually survived. They concluded that no method is reliable for predicting the mortality of surgical ICU patients. This illustrates the confusion that exists between interpreting an estimated probability of mortality and predicting whether a given patient will live or die. A good severity system provides an accurate estimate of the number of patients predicted to die among a group of similar patients; however, it does not provide a prediction of which particular patients will in fact die. Using a well-calibrated severity model, we can reasonably expect that approx. 75% of patients with a probability of mortality of 0.75 will die, but we cannot know in advance which of those patients will be among the 25% who will live. Furthermore, these 25% will not have falsified the odds but will have confirmed the validity of the probabilities.

The possibility that clinical decisions can be augmented by having an objective (although not always more accurate) assessment of a patient’s severity of illness is appealing. Physicians are interested in severity systems for individual patients as an adjunct to their informed but subjective opinion. Using these tools as part of the decision-making process is reasonable and prudent. Using these tools to dictate individual patient decisions is not appropriate. Decisions will and should remain the responsibility of the individual physician and should be based on a number of criteria, one of which is severity as estimated by a well calibrated scoring system.

Evaluation of ICU performance

Using the APACHE II system Knaus et al. [21] calculated the probabilities of hospital mortality in a sample of 16,622 consecutive patients from 42 ICUs and compared this to the actual outcome. They observed that the ratio of observed to predicted number of deaths varied from 0.67 to 1.21 across ICUs. That is, in some ICUs the observed mortality was lower than predicted by the models, and in some it was higher. Similarly, using the SAPS II system Le Gall et al. [22] compared the probabilities of hospital mortality and actual outcome in ICUs in several countries. They found that the ratio varied across countries from 0.74 to 1.31, with some countries having a lower number of deaths that predicted and some a higher number.

One cannot conclude from these findings, however that clinical performance in different ICUs or different countries is necessarily below par when the observed mortality is higher than predicted, or that it is necessarily above par when the observed mortality is lower than predicted. To use these ratios effectively one must know the extent to which they are affected by factors others than clinical performance. These ratios are most effectively interpreted as indicators that one should look more deeply into the situation in the various ICUs to identify factors associated with the observed mortality differential. These probabilities by themselves do not effectively control for all of the differences that may have an impact on outcome. They cannot control for differences in patient mix or for disparities in available technical and therapeutic resources. Neither can they control for administrative differences or the level or organization of support staffing (e.g., beds per nurse). Only after taking such factors into consideration can meaningful evaluations and comparisons be made.

Therapeutic trials

As a specific case in point, this discussion is oriented toward therapeutic trials for sepsis, but the issues involved can be applied to clinical trials involving any disease or condition and any proposed new therapy. While some authors [23] have stressed the importance of preexisting comorbidity for prognosis of septicemia in critically ill patients, others [24] have shown by multivariate analysis that using the initial score, cause (urosepsis or other), and treatment location prior to ICU admission provides the greatest degree of discrimination (ROC=0.82) of patients by risk of hospital death.

A complex model has been published for sepsis derived from a large database using physiology, primary disease, previous intensive care, age, clinical history of cirrhosis, and other variables [25]. This is proposed for use in clinical trials in which sepsis is the sole disorder of interest. However, the database from which the model was developed defined disease spectrum and inclusion criteria in a manner that may differ from that specified for a proposed trial of a new therapy for sepsis. In general it is unlikely that the precise inclusion or exclusion criteria for a specific trial were in used in compiling the original database from which a model was developed. Nor is it reasonable to expect that a large, general medical/surgical database would contain all of the information for addressing all the requirements of current and future trials. Although this should not deter one from the use of such models, it should make investigators wary of comparisons between the predicted mortality rate given by a model derived from a large database and the observed mortality rate in a precisely defined group. The probability can be used to stratify patients by level of severity at the onset of the trial, but conclusions about observed and predicted outcome should be drawn with care.

In a critique of scoring systems Loirat [26] suggested using a simpler tool without assigned weights for acute diseases. Such a disease-independent assessment of severity could be used to derive a disease-specific model using one-half of the patients in a control group. The model would be applied to the patients in the other one-half of the control group and the patients in the treatment group, and comparisons of observed and predicted outcome between the two groups could be made to evaluate the success of the treatment.

It must also be noted that the present general models have all been developed for use at very specific time periods, either at admission to the ICU (MPM_0), during the first 24 h of the ICU stay (SAPS II, APACHE III), or at three 24-h time points of the ICU stay (MPM_{24}, MPM_{48}, MPM_{72}). These models are not automatically transferable for use in stratifying patients at time of randomization in a clinical trial if this time point lies outside the time limits during which the models were intended to be applied. Research is necessary to confirm that severity at the time of randomization is accurately measured by these models (i.e., to confirm that they are well calibrated at the intended time period).

Conclusion

In an editorial Selker [27] stated that the desirable characteristics of risk-adjusted mortality predictors are that they be time-insensitive predictive instruments, based on the first minutes of hospital presentation, not affected by whether a patient is hospitalized, based on data collected in the usual care of patients, calibrated with a high degree of precision, integrated into computer systems, independent of the diagnosis-related groups system, and open for inspection and testing. These criteria are probably utopian, and the ideal scoring systems remains to be discovered. The available ICU scoring systems reviewed in this article are, however, based on rigorous research and have reported excellent calibration and discrimination.

Regarding the critical point of view we wish to stress the following: SAPS 3 seems very promising. It is currently the most recent and sophisticated model. The original models may be used to score patients' severity and make comparison of severity over years. The customized models are easy to obtain from the existing databases. The expanded SAPS II, simple to obtain from the existing data bases, may be used to compare performances over time.

References

1. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13:818–829
2. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A (1991) The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100:1619–1636
3. Le Gall J-R, Lemeshow S, Saulnier F (1994) A new Simplified Acute Physiology Score (SAPS II) based in European/North American multicenter study. JAMA 270:2957–2963

4. Lemeshow S, Klar J, Teres D, Avrunin JS, Gehlbach SH, Rapoport J, Rue (1994) Mortality probability models for patients in the intensive care unit for 48 or 72 hours: a prospective multicenter study. Crit Care Med 22:1351–1358
5. Aergerter P, Boumendil A, Retbi A, Minvielle E, Dervaux B, Guidet B (2005) SAPS II revisited. Intensive Care Med 31:416–423
6. Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B, Gouzes C, Lepage E, Moine P, Villers D (2005) Mortality prediction using the SAPS II: an update for French ICUs. Critical Care Med 9:R645–R652
7. Metnitz PGH, Moreno RP, Almeida E, Jordan B, Bauer P, Abizanda-Campos R, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR on behalf of the SAPS 3 investigators (2005) SAPS 3 – From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. Intensive Care Med 31:1336–1344
8. Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Abizanda-Campos R, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR on behalf of the SAPS 3 investigators (2005) SAPS 3 – From evaluation of the patient to evaluation of the intensive care unit. Part 2. Objectives, methods and cohort description. Intensive Care Med 31:1345–1355
9. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) Prognosis in acute organ-system failure. Arch Surg 202:685–693
10. Fagon JY, Chastre J, Novara A, Medioni P, Gibert C (1993) Characterization of intensive care unit patients using a model based on the presence or absence of organ dysfunctions and/or infection: the ODIN model. Intensive Care Med 19:137–144
11. Vincent JL, de Mendonca A, Cantraine F, Moreno R, Takala J, Suter PM, Sprung CL, Colardyn F, Blecher S (1998) Use of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Crit Care Med 26:1793–1890
12. Marshall J, Cook DJ, Christou NU, Bernard GR, Sprung CL, Sibbald WJ (1995) Multiple organ dysfunction score: a reliable description of a complex clinical outcome. Crit Care Med 23:1638–1652
13. Le Gall J-R, Klar J, Lemeshow S, Saulnier F, Alberti C (1996) The logistic organ dysfunction system. A new way to assess organ dysfunction in the intensive care unit. JAMA 276:802–810
14. Timsit JF, Fosse JP, Troché G, De Lassence A, Alberti C, Garrouste-Orgeas M, Azoulay E, Chevret S, Moine P, Cohen Y (2001) Accuracy of a composite score using daily SAPS II and LOD scores for predicting hospital mortality in ICU patients hospitalized for more than 72 h. Intensive Care Med 27:1012–1021
15. Lemeshow S, Le Gall JR (1994) Modeling the severity of illness of ICU patients. JAMA 272:1049–1055
16. Hosmer DW, Lemeshow S (1989) Applied logistic regression. Wiley, New York
17. Hanley JA, Mc Neil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
18. Le Gall J-R, Lemeshow S, Leleu G, Klar J, Huillard J, Montserra R, Teres D, Artigas A for Intensive Care Unit Scoring Group (1995) Customized probability models for early severe sepsis adult intensive care patients. JAMA 273:644–650
19. Teres D, Lemeshow S (1994) Why severity models should be used with caution. Crit Care Clin 10:93–110
20. Meyer AA, Messick WJ, Young R, Backer CC, Fakhry S, Muakkassa F, Rutherford EJ, Napolitano LM, Rutledge R (1992) Prospective comparison of clinical judgement and APACHE II score in predicting the outcome in critically ill surgical patients. J Trauma 32:747–753
21. Knaus WA, Wagner DP, Zimmerman JE, Draper EA (1993) Variation in mortality and length of stay in intensive care units. Ann Intern Med 118:753–761
22. Le Gall J-R, Artigas A, Lemeshow S, Saulnier F, Avrunin J (1993) Une comparaison internationale des unités de réanimation (abstract). Reanimation Soins Intensifs Med Urgence 6:656
23. Pittet D, Thievent B, Wenzel RC, Gurman G, Sutter PM (1993) Importance of preexisting comorbidities for prognosis of septicemia in critically ill patients. Intensive Care Med 19:265–272
24. Knaus WA, Sun X, Nystrom Pr, Wagner DP (1992) Evaluation of definitions for sepsis. Chest 101:1656–1662
25. Knaus WA, Harrel FE, Fisher CJ Jr, Wagner DP, Opal SM, Sadoff JC, Draper EA, Walawander CA, Conboy K, Grasela TH (1993) The clinical evaluation of new drugs for sepsis: a prospective study design based on survival analysis. JAMA 270:1233–1241
26. Loirat P (1994) Critique of existing scoring systems: admission scores. Reanation Soins Intensifs Med Urgence 3:173–175
27. Selker HP (1993) Systems for comparing actual and predicted mortality rates: characteristics to promote cooperation in improving hospital care. Ann Intern Med 118:820–822