Intensive Care Med (1995) 21:356–360

BRIEF REPORT

E. Féry-Lemonnier
P. Landais
P. Loirat
D. Kleinknecht
F. Brivet

Evaluation of severity scoring systems in ICUs – translation, conversion and definition ambiguities as a source of inter-observer variability in Apache II, SAPS and OSF

Received: 16 November 1993
Accepted: 30 May 1994

E. Féry-Lemonnier · P. Landais (✉)
Laboratoire de Biostatistique
et d'Informatique Médicale,
Hôpital Necker-Enfants Malades,
149, rue de Sèvres, F-75743, Paris, France

P. Loirat
Hôpital Foch, Suresnes, France

D. Kleinknecht
Hôpital de Montreuil, Montreuil, France

F. Brivet
Hôpital A. Beclère, Clamart, France

Abstract *Objective:* To explore translation, conversion and definition ambiguities, when using severity scoring systems in patients admitted to intensive care units (ICUs).
Design: A prospective study of the prognosis of acute renal failure in ICUs.
Setting: The study was conducted in 20 French ICUs.
Patients: 360 patients presenting with severe acute renal failure were studied during their ICU stay.
Measurements and results: The inter-observer variability of Apache II (acute physiology and chronic health evaluation), SAPS (simplified acute physiology score), and OSF (organ-system failure) was considered. For Apache II, we explored the uncertainty of measurements arising from conversion into SI units, the rounding procedures used for the non-inclusive intervals defined for quantitative parameters such as age, mean arterial pressure (MAP) or serum creatinine, the absence of definition of acute renal failure (ARF) and its consequence on doubling serum creatinine values, and the absence of guidelines in the case of spontaneous ventilation when arterial blood gases (ABG) and forced inspiratory oxygen (FIO_2) were not measured. The resulting variability was evaluated, calculating the lowest and the highest value of the scoring system for each patient. The mean difference by patient was greater than 1.5 ($p < 0.0001$). Other examples were presented and discussed for SAPS and OSF.
Conclusions: Translation, conversion and definition ambiguities are a source of inter-observer variability and increase the risk of classification and/or selection biases. This gives rise to particular concern in the design and analysis of multicenter trials or meta-analysis, and improvement of these scoring systems should be envisaged in the future.

Key words Evaluation · Scoring systems · Inter-observer variability · Acute renal failure

Introduction

Since the early 1980s, evaluation in ICUs has been directed towards new technologies, new therapeutic strategies, evaluation of activities and services, cost analysis [1], and clinical decision making. The most important goal of ICU activity is to decrease mortality. Identification of patients with a high risk of mortality in ICUs is generally based on the use of scoring systems. Patients are classified according to prognostic factors defined by clinical experts or by specific statistical analysis. The higher the score on admission, the higher the risk of mortality. This predictive approach is supposed to improve the choice of

diagnostic tests and treatments, and according to some authors, to enable decisions to be made in accordance with the expected severity of the critically ill patient [2, 3]. Estimation of the expected evolution of patients admitted to ICUs is also possible, although individual prediction remains controversial [4, 5]. Finally, it allows clinical trials to be planned using prognositc factors as criteria for inclusion, stratification or adjustment, and is thus of major interest for multicenter studies with patient recruitment varying from one ICU to another.

The main prognostic factors are age, health status prior to admission to the ICU, and factors related either to the acute status of the disease leading to ICU admission or to the severity of the resulting physiological disorders [6–9]. The scoring systems consist of several of the above prognostic factors, with a mortality probability scale.

Apache II, SAPS and OSF were studied in 360 patients admitted to an ICU and presenting with ARF [10, 11]. Prognostic values of the scores were analyzed using receiver operating curves (ROC) and logistic regression, the results of which are presented elsewhere [10, 11]. The difficulties encountered in applying these scoring systems, which are widely used in France, are described herein. These difficulties were related to the looseness of several definitions, together with ambiguities arising from translation into French, and unexpected problems of conversion from traditional units to SI units.

Materials and methods

A prospective study was conducted over a 6-month period in 20 multidisciplinary ICUs to assess the prognosis of patients hospitalized with severe ARF. Inclusion criteria were blood urea greater than 36 mmol/l and/or serum creatinine levels greater than 310 μmol/l [4]. In patients with previous chronic renal insufficiency (serum creatinine levels between 150 and 300 μmol/l), the criterion for ARF was a 100% increase in blood urea and/or serum creatinine levels above baseline values. Physiological variables were measured on ICU admission, on inclusion (when ARF occurred during the ICU stay), and on days 2, 4, and 7. The raw data were taken from the medical questionnaires, and Apache II, SAPS, and OSF were then calculated automatically.

The Apache II physiological score included 12 items [12]. The weighting system was based on a scale of 0–4. Sixteen biological variables were in fact necessary so as to calculate the Apache II score, and FIO_2, P_aO_2, and P_aCO_2 were required in order to calculate oxygenation. Fifteen variables were required in order to calculate the SAPS score on the basis of 14 items, weighted from 0–4 [13, 14]. The five OSFs were defined by 15 items [4, 15], and 17 physiological criteria had to be collected for the diagnosis of OSF. Three criteria (FIO_2, P_aCO_2, P_aO_2 were necessary for calculation of A_aDO_2. Results are expressed as mean ± SD for quantitative variables. A paired *t*-test was used to evaluate intragroup variability.

Results

Three hundred and sixty patients aged 60 ± 18 years (240 male and 120 female) were enrolled in the study. ARF was present in 217 cases on admission (death rate = 49.7%), and occurred in 143 patients during the ICU stay (death rate = 71.3%). The mean scores on admission were: SAPS 16.8 ± 0.3, Apache II 24.4 ± 0.4, and OSF 1.47 ± 0.05. The descriptive and analytical results are described elsewhere [10, 11]. The presentation of the results reflects the inter-observer variability arising when establishing the scores.

Apache II

Age is not an integer when calculated from the date of birth. For a patient aged 44.5 years, either 0 or 2 points might be added to the score, depending upon the choice of the physician as presented in Table 1.

After calculation, several variables are obtained with one or two decimals (MAP, P_aO_2 or A_aDO_2). However, the limits of each Apache II category are integers. An example is given in terms of MAP in Table 1. An MAP of 69.5 mmHg does not correspond to any predefined category for this variable; it may be rounded to either 69 or 70. We interpreted the limit values in terms of strict non-equivalence: for example, the class of MAP ≥ 50 mmHg and ≤ 69 mmHg, was changed to MAP ≥ 50 mmHg and < 70 mmHg. However, rounding to the upper unit rather than to the lower leads to a variation of 1 or 2 points per variable, and of several points per patient.

SI units are widely used in France. Serum creatinine concentration is given in μmol/l. For Apache II, serum creatinine is given in mg/dl, 1 mg/dl corresponding to 88.4 μmol/l and 1.13 mg/dl to 100 μmol/l. For classification purposes we could choose to convert the values of serum creatinine obtained into mg/dl, or the original classification into SI units (Table 2). If we convert the values of serum creatinine into SI units for a range of 125–132 μmol/l, the resulting values, ranging between 1.41 and 1.49 mg/dl, do not definitively-fit in either the third or the fourth category of Table 2. Conversely, when the original limits are converted into SI units, it can be noted that there are gaps between the categories. Thus, a serum creatinine value of 125 μmol/l does not fit into any category. We defined new limits in accordance with our experts; however, generally accepted limits are required in SI units so as to avoid such classification ambiguities.

Table 1 Apache II, limits and decimals: age and mean arterial pressure (*MAP*)

Points	0	+2	+3	+4	+5	+6
Age (years)	≤44	45–54	55–64	–	65–74	≥75
MAP (mmHg)	109–70	69–50	–	≤49	–	–

Table 2 Apache II, limits and conversion into SI units: serum creatinine values

Limits / Points	+4	+3	+2	0	+2
Knaus original (mg/dl)	≥3.5	3.4–2	1.9–1.5	1.4–0.6	<0.6
Knaus converted (μmol/l)	≥309	301–177	168–133	124–53	<53
Bedock [1] (μmol/l)	≥318	317–180	179–136	135–54	<54
Present work (μmol/l)	≥309	<309≥177	<177≥133	<133≥53	<53

Serum creatinine scoring doubles in patients presenting with ARF. However, ARF definition is not generally recommended nor is it specified in Knaus' classification system. We chose the pragmatic definition provided by our protocol.

Measurement of an 12 physiological values is mandatory when using Apache II. Missing data results in the patient record being excluded. ABG were not measured in 14 patients. According to clinical judgement, we considered that the results for patients breathing spontaneously ventilation should be in the normal range (7 cases); this choice is clearly physician dependent. FIO_2 was not recorded in 71 patients. No guidelines are provided for this parameter particularly as regards spontaneous ventilation. We allocated a value of $FIO_2 < 50\%$ to all patients under spontaneous ventilation (59 cases).

The variability of Apache II arising from these inaccurate results was evaluated by calculating two scores for each patient. One was calculated from the lowest value of the three variables presented above (MAP, serum creatinine, definition of ARF) and the other, from the highest values. The mean difference by patient between the lowest and highest values was greater than 1.5 points ($p < 0.0001$). The impact of these inaccurate results on Apache II is not therefore negligible.

Assignment of chronic health points depends on the circumstances of patient admission (non-operative or emergency postoperative patients versus elective postoperative patients). A history of severe OSF or an immunocompromised state requires approximately 15 additional criteria to be assessed. Several of these variables are unclear (secondary polycythemia, recent high dose corticotherapy, etc.) and are not easy to characterize a posteriori.

Simplified acute physiological score

No written guidelines are provided for missing data. In the original article, only 50% of patient records were complete and 3 variables maximum were missing from each of the incomplete patient records [13].

The Glasgow Coma Score is defined in the absence of sedation; nevertheless, sedation policy for ventilated patients may vary from one center to another; hence a, Glasgow Coma Score may well not be given. In our study, 30 patients of whom 28 were ventilated had a Glasgow Coma Score lower than 6 on admission. Thus, when the score is not evaluated at the patient's bedside, it is advisable to assess an additional variable indicating sedation in ventilated patients.

In the case of cardiac arrest four points are added to heart rate, systolic blood pressure, respiratory rate and the Glasgow Coma Score [14]. It is difficult to identify cardiac arrest a posteriori or on the basis of raw data. Blood pressure equal to zero does not necessarily correspond to cardiac arrest since most sphygmomanometers lack the precision necessary to record blood pressure levels lower than 20 mmHg, and respiratory rate is of little use if a patient is ventilated. We therefore suggest the creation of a specific item, namely cardiac arrest, if SAPS in not recorded at the patients bedside.

Diuresis had to be extrapolated to a 24-h period for patients staying in an ICU for less than 24 h [14]. As regards ARF, the results might then be biased if this parameter is considered a factor of disease severity.

Organ-system failure (OSF)

In the original article defining OSF, no guidelines are provided for missing data [4]. There were no missing values in our group; however, if there were, it seems that the patients should be excluded from analysis. We would presume that both Apache II and OSF would adopt a similar policy regarding missing data since they have the same authors. However, not being clearly stipulated, but is left to the physician to decide, thus increasing variability.

Day 1 of OSF is not the first day of ICU stay, but that of the first OSF. Hence, the date of admission to the ICU does not necessarily represent day 1 of OSF. The origin of OSF is not fixed but variable on the temporal axis; hence, the physiological variables need to be assessed on a daily basis, even before the occurrence of the first OSF. In the original paper only 38% of patients presented with one or more OSF on admission [4].

The definition of respiratory failure seemed ambiguous. One of the criteria for respiratory OSF is the presence of assisted ventilation 72 h after the onset of OSF (or 24 h according to a recent article [15]). It is not specified whether this refers to the 72 h after the onset of respiratory. OSF, defined by one of the other criteria, or after the

onset of another OSF. Knaus et al. (text to Table 4 [4]) states that the patient "was dependent on a ventilator after the initial 3 days (72 h) in the ICU." Thus, whether or not OSF is recorded (highly prevalent in our group, 52% of patients being ventilated on admission) might depend on the interpretation of the original criteria. In our group, when we considered the first 24 h in the ICU, we could not attribute respiratory OSF in every case of assisted ventilation, despite the evidence of multiple-organ failure in these ARF patients.

In the article by Rauss et al. [15], MAP was replaced by systolic blood pressure (SBP), without taking diastolic blood pressure (DBP) into account. However, in our group DBP was a significant prognostic variable of hospital mortality ($p < 10^{-4}$) using a logistic regression model. Fifty patients in our group fulfilled the American criterion (i.e., MAP $\leq$ 49 mmHg), as opposed to only 24 patients who fulfilled the French criterion (SBP $<$ 60 mmHg). Finally, as stated earlier, the definition of respiratory failure is different in the American and the French articles, the 4th day and the 2nd day of OSF being chosen, respectively [4, 15].

Discussion

Severity scoring systems are now widely used in intensive care medicine. Many publications have aimed to design, improve and analyze the contribution of severity scoring systems in intensive care practice. However, few have focused on the problems raised by translation of severity scoring systems into foreign languages. We have drawn attention to several practical problems concerning this subject, that arose during a multicenter study on ARF in ICUs conducted in France. The problems encountered when using Apache II, SAPS and OSF were loose definition of medical terms, and translation and classification ambiguities arising from conversion into SI units, rounding procedures, missing data or procedures for grading clinical abnormalities.

Appropriate assigment of ARF or chronic health points is difficult since standardized definitions are not clearly stated. Standardization of medical terminology is crucial, and particularly so for international classifications. The lack of definition with regard to medical terminology also affects international classifications such as SNOMED (*Systematized Nomenclature of Medicine*) used by clinicians, or MeSH (Medical Subject Headings) used for classification of medical literature. Unambiguous definitions of syndromes or diseases are necessary so as to avoid classification biases.

When the values of decimals and limits are not strictly defined, rounding procedures become necessary. Variation in the values of the score is thus introduced, illustrated by two examples concerning age and MAP for Apache II. The consequence of choosing the lower limit resulted in allocating two points to these items, as opposed to zero points when the upper limit was considered. Such variability is physician-dependent, and decision making varies from one observer to another. A decision-making process common to all physicians, is thus necessary.

Conversion procedures give rise to a similar problem: the conversion of Apache II serum creatinine limits into the SI units used in France may result in different values being obtained for the converted limits. Limits should therefore be defined in SI units for international use.

Assignment of normal results in the case of missing data is clearly physician-dependent, as we demonstrated for the missing ABG results. In calculation of each patient's score, the personal nature of the physician's decision must be borne in mind. We showed that the impact of these decisions for Apache II may result in significant score variability. The time of onset of OSF is not identical to the beginning of Apache II or SAPS, since the date of admission to the ICU does not necessarily represent day 1 of OSF. Day 1 of OSF is the day the first OSF occurred, which thus varies in terms of time. This must be kept in mind when comparing OSF with Apache II or SAPS, since the last two scoring systems are validated on admission only.

These examples, taken from our own experience, outline potential biases encountered when using severity scoring systems. It appeared that definition, translation and conversion ambiguities are potential sources of inter-observer variability, and may be of important concern in the design and analysis of multicenter trials or meta-analysis. Further recommendations are necessary to reduce inter-observer variability and the risk of erroneous classification.

Acknowledgements We thank the physicians, members of the French Study Group on ARF. Participating centres: Annonay: B. Bedock; Argenteuil: G. Bleichner; Bicêtre: R. Sananes, C. Richard; Clamart: J. Dormont; Colombes: F. Coste; Créteil: C. Brun-Buisson; Eaubonne: R. Moreau, C. Sicot; Evry: D. Perrin-Gachadoat, A. Tenaillon; Garches: P. Gajdos; Gonesse: F. Blin; Lille: F. Saulnier; Montreuil: M.M. Agostini; Nantes: F. Nicolas; Paris Saint-Antoine: F. Staïkowsky, G. Offenstadt; Paris Saint-Joseph: J. Carlet; Saint-Germain: G. Perseil, I.L. Ricome; Saint-Brieuc: G. Guivarch; Saint-Denis: B. Verdière; Strasbourg: I. Runge, J.D. Tempe; Suresnes: P. Mezzarobba. This work was supported by the Caisse Régionale d'Assurance Maladie de l'Ile de France (CRAMIF), the Institut de la Santé et de la Recherche Médicale (INSERM) and the Université René Descartes, Paris V, France

References

1. Le Gall JR, Loirat P (1990) Evaluation en réanimation. (Collection d'anesthésiologie et de réanimation) Masson, Paris
2. Chang RWS, Jacobs S (1986) Use of Apache II severity of disease classification to identify intensive-care-unit patients who would not benefit from total parenteral nutrition. Lancet II: 1483–1487
3. Wagner DP, Knaus WA, Draper EA, Zimmerman JE (1983) Identification of low-risk monitor patients within a medical-surgical intensive care unit. Med Care 21:425–434
4. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) Prognosis in acute organ-system failure. Ann Surg 202:685–693
5. Chang RWS (1989) Individual outcome prediction models for intensive care units. Lancet II:143–146
6. Le Gall JR (1986) The intensive care unit: definitions and managerial differences. A French multicentric study on 38 units. Elsevier, Amsterdam, pp 39–54
7. Le Gall JR, Brun-Buisson C, Trunet P, Latournerie J, Chantereau S, Rapin M (1982) Influence of age, previous health status, and severity of acute illness on outcome from intensive care. Crit Care Med 10:575–577
8. French multicenter group of ICU research; INSERM Unit 169 of statistical and epidemiological studies (1989) Factors related to outcome in intensive care: French multicenter study. Crit Care Med 17:305–308
9. Knaus WA, Wagner DP, Draper EA et al (1991) The Apache III prognostic system. Chest 100:1619–1636
10. The French Study Group on ARF, Agostini MM, Mezzarobba P, Kleinknecht D, Loirat P, Brivet F, Landais P (1992) Prognosis of acute renal failure in intensive care units. A prospective multicenter study. Intensive Care Med 18 [Suppl 2]:S65
11. French Multicenter Group, Mezzarobba P, Agostini MM, Loirat P, Kleinknecht D, Brivet F, Landais P (1992) Comparison of SAPS, Apache II and OSF in the evaluation of severe acute renal failure. Intensive Care Med 18 [Suppl 2]:S65
12. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) Apache II: a severity of disease classification system. Crit Care Med 13:818–829
13. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, Mercier P, Thomas R, Villers D (1984) A simplified acute physiology score for ICU patients. Crit Care Med 12: 975–977
14. Le Gall JR, Loirat P, Alperovitch A (1989) The simplified acute physiology score (SAPS). Probl Crit Care 3: 578–584
15. Rauss A, Knaus WA, Patois E, Le Gall JR, Loirat P (1990) Prognosis from multiple organ system failure: the reliability of objective estimates for survival. Med Decis Making 10:155–162