Intensive Care Med (1999) 25: 140–142
© Springer-Verlag 1999

D. Teres
S. Lemeshow

When to customize a severity model

Accepted: 20 November 1998

D. Teres (✉)
Center for Health Services Research, Baystate Medical Center,
759 Chestnut St., Springfield, MA 01199, USA
email: daniel.teres@bhs.org
Tel. + 1-(413)-784-9778
Fax + 1-(413)-784-3103

S. Lemeshow
University of Massachusetts,
School of Public Health and Health Sciences, Amherst,
MA 01003, USA

General intensive care unit (ICU) severity models are remarkably popular, as judged by the number of peer review publications that contain a severity system as a key component of the study. A MEDLINE search of articles published between January 1993 and December 1997 revealed that there were 552 articles published, with two-thirds of these studies using the Acute Physiology and Chronic Health Evaluation (APACHE), mortality probability models (MPMs), or the Simplified Acute Physiology Score (SAPS) (S. Weitzen, T. Higgins, D. Teres, unpublished data). Studies using pediatric or neonatal models were next in frequency (15 %), with a much smaller number of articles focused on cardiac surgery models, multiple organ failure models, or trauma scores. In more than half of these studies the severity score was used for risk stratification or as a clinical descriptor of patient populations. Surprisingly, 30 % of the articles focused on the development, validation, or performance of severity systems or comparisons of models. Other studies (10 %) used models as part of economic analyses and in only 4 % of studies was the primary goal comparison of quality of care in ICUs.

In this issue of *Intensive Care Medicine*, Metnitz et al. [1] report, in a small but detailed study, on the performance of SAPS II in nine Austrian ICUs. The results are not surprising; SAPS II did not fit well in the new set of patients 6 years after the original publication of SAPS II [2]. To date, there have been no studies that have shown that severity models are stable over time, in a new setting, and with different case mixes [3–7]. APACHE III did not demonstrate external validation using a large number of hospitals in the United States [8, 9]. If we believe in medical progress and in the advancement of science, then we would not expect the models to show good calibration over time. We would expect the observed mortality to be lower than the mean predicted hospital mortality, particularly for middle severity patients. If there were severe resource constraints and reduction in staffing and we had good benchmark data, we might expect to find higher observed than predicted hospital mortality. What about the introduction of a new disease now more frequently treated in the ICU but not previously included in severity model development? AIDS was considered a fatal disease and was only minimally included in APACHE, MPM, and SAPS databases [2, 10–12]. Now AIDS is a more chronic disease with more patients being admitted to the ICU with AIDS as a background condition or as a major component of the acute process. What about new technology? Non-invasive positive pressure ventilation is now being more commonly used. For a patient with acute exacerbation of chronic respiratory failure who is placed on nocturnal nasal ventilation on admission to the ICU, how do we apply the definitions of respiratory failure including intubation and mechanical ventilation as described in models that are now several years old? For all of the above reasons, there is a rationale for considering a role for recalibrating or customizing a severity model for a specific application.

What are the advantages of such an approach? When should this mathematical manipulation be done? Are there any alternative approaches? What information is lost when a model is customized or recalibrated?

The first article on the practical aspects of customizing ICU severity models was stimulated by the need to have accurate severity models available for patients with sepsis who were being enrolled in phase II/III clinical trials [13]. The alternative approach would have been to launch a new, large data collection effort focused on patients admitted to the ICU with clearly defined sepsis (Systemic Inflammatory Response Syndrome with severe sepsis) and then to develop a unique model for this important subgroup of ICU patients. Unfortunately, developing a new model de novo is time consuming and an expensive endeavor. The approach devised by Le Gall and Lemeshow was to customize MPM and SAPS for patients with sepsis by mathematically adjusting the logit so that, in the revised model, there was good correspondence between observed and predicted mortality [13]. In clinical trials it is important to have properly calibrated models for risk stratification and for measuring severity-adjusted efficacy [14, 15]. Zhu et al. [16] published a detailed computer simulation study using the MPM II database. The primary goal of this study was to provide a statistical basis for using severity models for comparative quality assessment. They evaluated the impact of various mortality rates by arbitrarily and systematically changing the outcome at hospital discharge from "survived" to "deceased" or from "deceased" to "survived." They also systematically changed the sample size. They showed that goodness-of-fit calibration was more sensitive to differences in mortality than discrimination as measured by the receiver operator characteristic curve. The sample size also had an effect on model "stability." They concluded that severity models were useful for quality assurance purposes. To accommodate likely improvements in ICU technology, they demonstrated two techniques for recalibrating a model through customizing either the logit or individual coefficients [16].

What followed is an example of the law of unintended consequences. There have now been a number of studies in the literature that have evaluated model performance with very adequate attention to data collection and data management. When the goodness-of-fit test was applied and the model was demonstrated to be poorly calibrated, the researchers then assumed that the next step was to customize the model [17, 18]. The presumption was that there was something wrong with the model which either under- or overpredicted outcome of vital status at hospital discharge. Quality of care was considered to be a less likely explanation. Case mix was "corrected" by recalibration.

There are some circumstances where customization is appropriate. In the Greater Cleveland Quality Choice study, the authors recalibrated APACHE III prior to initiating their quality of care comparison [19]. They then proceeded to perform a quality of care report card study without further recalibrating the model. Once there is a good, up-to-date local model, that severity system can be used for quality of care comparisons, assuming high quality data collection techniques, clarification of definitions, and good data management. The Austrian study reported in this issue provides good information about the attention paid to these details, including having one expert data collector travel to each of the sites [1]. Hopefully, the Austrian research group will now proceed with a quality of care comparison. The analysis should determine whether each ICU is above, below, or at the referent point set by the severity model. There should not be a rationale to recustomize the model but rather to focus on explaining the differences based on case mix and/or quality of care differences.

Our concern is that, in the decision to customize the model to improve model performance, important information may be lost. In Tables 2 and 3 and in Fig. 1 a of the Metnitz et al. study, there is clear demonstration that the observed mortality is lower than expected for patients in the middle severity strata [1]. This finding does not appear to be random and should be viewed as positive. Perhaps care has improved over the 6-year time period! In the studies by Apolone et al., there was a striking geographic difference [5]. Northern Italy had a standardized mortality ratio that was similar to that published in the original SAPS II paper, while central and southern Italy had a higher observed mortality than predicted. Is this merely "underprediction" by a bad or poorly calibrated model which could be fixed by customizing the model, or is there a quality of care difference? In the EURICUS database there was wider variability in the observed to expected mortality ratio, indicating the difficult issues related to case mix differences [20].

In the United States, there is a trend toward lowered observed compared to predicted hospital mortality. The reason is that, because of managed care pressure, some ICU patients with complex care issues and ventilator dependence (DRG 438) are being sent to a separate facility for continued care in a subacute or postacute care facility. These ICU patients are discharged from the acute care hospital and, for severity of model application, they are considered discharged alive. However, some of these patients may be subsequently admitted to another acute care hospital or have further complications and die in the subacute facility.

To accommodate this major change in medical practice and to better define when critical care starts, our research group has proposed focusing attention on the acute episode of critical illness rather than on each ICU admission [21]. A reasonable endpoint would be vital status at 90 days. For such an approach, the severity model will need to be recalibrated for 90 days instead of hospital discharge. Hopefully, data bases will adjust to the proposed change and we can then compare ICU and cost performance from the beginning of the acute critical episode to a more acceptable endpoint using previously defined methods [22]. As with any proposed change, there are potential problems. These include defining when the first and only severity measure should be collected and how to track patients over the 90 days. ICU severity models have been successful because the clock starts when you first see the patient in the ICU and stops when the patient leaves your hospital. Despite the new problems to be encountered, the time has come to customize the severity models for the acute episode of critical illness.

References

1. Metnitz PGH, Valentin A, Vesely H, Alberti C, Lang T, Lenz K, Steltzer H, Hiesmayr M (1999) Prognostic performance and customization of the SAPS II: results of a multicenter Austrian study. Intensive Care Med 25: 192–197
2. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPS II) based on European/North American multicenter study. JAMA 270: 2957–2963
3. Nouira S, Belghith M, Elatrous S, Jaafoura M, Ellouzi M, Boujdaria R, Gahbiche M, Bouchoucha S, Abroug F (1998) Predictive value of severity scoring systems: comparison multicenter study of four models in Tunisian adult ICUs. Crit Care Med 26: 852–859
4. Beck DH, Taylor GL, Millar B, Smith GB (1997) Prediction of outcome from intensive care: a prospective cohort study comparing Acute Physiology and Chronic Health Evaluation II and III prognostic systems in a United Kingdom intensive care unit. Crit Care Med 25: 9–15
5. Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G, Melotti RM (1996) The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Intensive Care Med 22: 1368–1378
6. Bastos PG, Sun X, Wagner DP, Knaus WA, Zimmerman JE (1996) The Brazil APACHE III Study Group: application of the APACHE III prognostic system in Brazilian intensive care units: a prospective multicenter study. Intensive Care Med 22: 564–570
7. Moreno R, Reis Miranda D, Fidler V (1998) Evaluation of two outcome prediction models on an independent database. Crit Care Med 26: 50–61
8. Zimmerman JE, Wagner DP, Draper EA, Wright L, Alzola C, Knaus WA (1998) Evaluation of Acute Physiology and Chronic Health Evaluation (APACHE) III predictions of hospital mortality in an independent database. Crit Care Med 26: 1317–1326
9. Teres D, Lemeshow S (1998) As American as apple pie and APACHE. Crit Care Med 26: 1297–1298
10. Knaus WA, Draper EA, Wagner DP, Zimmerman JE (1985) APACHE II: a severity of disease classification system. Crit Care Med 13: 818–829
11. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A, Harrell FE Jr (1991) The APACHE III prognostic system: risk prediction of hospital mortality for critically ill hospitalized adults. Chest 100: 1619–1636
12. Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH, Rapoport J (1993) Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. JAMA 270: 2478–2486
13. Le Gall JR, Lemeshow S, Leleu G, Klar J, Huillard J, Rue M, Teres D, Artigas A (1995) Customized probability models for early severe sepsis in adult intensive care patients. *JAMA* 273: 644–650
14. (1997) From the bench to the bedside: The future of sepsis research. Executive summary of an American College of Chest Physicians, National Institute of Allergy and Infectious Disease, and National Heart, Lung, and Blood Institute workshop. Chest 111: 744–753
15. Lemeshow S, Teres D, Moseley S (1997) Statistical issues in clinical sepsis trials. In: Fein AM, Abraham E, Balk R, Bernard G, Bone RC, Dantzker D, Fink M (eds) Textbook of Sepsis and Multiorgan Failure. Williams & Wilkins, Baltimore, pp 614–626
16. Zhu B, Lemeshow S, Hosmer DW, Klar J, Avrunin JS, Teres D (1996) Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: A simulation study. Crit Care Med 24: 57–63
17. Moreno R, Apolone G (1997) Impact of different customization strategies in the performance of a general severity score. Crit Care Med 25: 2001–2008
18. Sarmiento X, Rue M, Guardiola JJ, Toboso JM, Soler M, Artigas A (1997) Assessment of the prognosis of coronary patients: performance and customization of generic severity indexes. Chest 111: 1666–1671
19. Sirio CA, Shepardson LB, Rotondi AJ, Cooper GS, Angus DC, Harper DL, Rosenthal GE (1999) Community-wide assessment of intensive care outcomes using a physiologically-based prognostic measure: Implications for critical care delivery from Cleveland Health Quality Choice. Chest (in press)
20. Moreno R, Apolone G, Reis Miranda D (1998) Evaluation of the uniformity of fit of general outcome prediction models. Intensive Care Med 24: 40–47
21. Teres D, Higgins T, Steingrub J, Loiacono L, McGee W, Circeo L, Brunton M, Giuliano K, Burns M, Le Gall JR, Artigas A, Strosberg M, Lemeshow S (1998) Defining a high performance ICU system for the 21st century: a position paper. J Intensive Care Med 13: 195–205
22. Rapoport J, Teres D, Lemeshow S, Gehlbach S (1994) A method for assessing the clinical performance and cost effectiveness of intensive care units: a multicenter inception cohort study. Crit Care Med 22: 1385–1391