EDITORIAL

J. Geoffrey Chase
Geoffrey M. Shaw

How standard is the "S" in SMR?

J. G. Chase (✉)
Department of Mechanical Engineering,
Centre for Bio-Engineering, University of Canterbury,
Private Bag 4800, Christchurch, New Zealand
e-mail: geoff.chase@canterbury.ac.nz
Tel.: +64-3-3642987
Fax: +64-3-3642078

G. M. Shaw
Department of Intensive Care Medicine, Christchurch Hospital,
University of Otago School of Medicine, Private Bag 4710,
Christchurch, New Zealand
e-mail: geoff.shaw@cdhb.govt.nz
Tel.: +64-3-3640640
Fax: +64-3-3640099

Prognostic models are both a useful everyday tool and an ongoing "holy grail" in many fields of medicine. Their potential and use in clinical management of patients can have significant health, social and economic impact [1–4]. Hence, the search for better, more accurate, and better discriminating models that require only the most certain and measurable of input data is an ongoing task [3, 5, 6].

A primary goal is to ensure these models are adequately validated in development. A second major goal in this search is to ensure utility of the prognostic model through the need for minimal or straightforward input data. Finally, it is important to understand the longevity of such models as they all must change with time, as cohorts, treatment approaches and clinical cultural changes all take a toll on their accuracy.

The work presented by Minne et al. [7] in this issue of *Intensive Care Medicine* meets all three main goals in presenting and analysing a SAPS-II-based prognostic model for mortality in elderly intensive care unit (ICU) patients. The work is quite unique in its presentation of a general analysis approach to these models. It presents a robust prospective validation. More importantly, it analyses the model over a 5-year period with respect to its predictive performance, including the effect of recalibration. As a result, it clearly defines the effective longevity of the model.

The critical outcome to our minds is the general and robust framework within which this analysis was performed. Equally, the approach taken is readily generalisable to any similar problem. Hence, a major part of the value of this work is in providing a robust, generalisable and relatively straightforward framework within which to validate and analyse prognostic models—a significant achievement that could lead towards standardizing (to an extent) the approach to such models in the field, as well as the expectations that the creators of such models would have to meet. These steps would be a major advance towards creating more trust and transparency in the development and use of these models.

As noted by the authors, the use of prognostic models for mortality has a long history in critical care. One main use has looked towards managing patient care and treatment decisions [4]. A greater use has been for delineating and comparing ICU performance or quality [8–10]. Their use is becoming more evident as critical care resources become increasingly stressed. However, it is toward this latter benchmarking use, wherein model predictions must be adjusted for severity of illness and thus expected mortality, that this work by Minne et al. [7] raises a potential intriguing and perhaps slightly mischievous point.

Specifically, the results presented are robust in noting discriminatory power, both with and without repeated recalibration over time, for multiple parameters. The first two parameters, area under the receiver operating characteristic curve (AUC) and the Brier score, are both well known and directly reflect the discriminatory power of the model. It is no surprise that a proven model with this robust analysis shows solid, expected results with regard to these metrics.

The third metric—standardized mortality rate (SMR)—is more interesting. Because SMR is adjusted on the basis of expected mortality for a cohort it thus reflects the confluence of observed and predicted mortality. However, the results in Minne et al. [7], as shown over time, raise an interesting point. While SAPS-II is relatively constant or rising over time, SMR without recalibration is falling steadily in the control chart results presented. Hence, only 15% of ICUs surveyed had SMRs greater than 1, and repeated recalibration only raised this value to 35%. A "perfect" model would potentially reflect a 50:50 split—raising questions about SMR as a surrogate of care.

One readily possible cause is simply that all units are improving the quality of care and that model predictions are outdated, but this issue would theoretically be managed by repeated recalibration. However, this behaviour does not match the results for the more certain and readily measured AUC or Brier score metrics, which were excellent. Hence, while there is no ambiguity in the determination of AUC or Brier score relative to a model, can SMR be affected in ways that models do not account for?

A potential note of interest is that during the prospective period of the study, the authors report that the numbers of beds and admissions had increased from prior periods in some units studied. In particular, when ICUs expand they allow greater access to patients who previously might have been cared for outside of ICU, e.g.

patients undergoing major vascular surgery. Specifically, prognostic models do not capture effects similar to lead time bias [11, 12].

In this scenario, under-resourced ICUs might only accommodate patients that demonstrated failure to respond to therapy. While non-responders may still have the same acute physiological derangements as they had 24 h before, they are now clearly different from cohorts of responders. Thus, if beds are scarce, non-responders are preferentially admitted and may well have a greater mortality compared with the broad model prediction, all else being equal.

Model-evaluated SMR may not adequately reflect this change without regular reconstruction and/or recalibration. Equivalently, all models only approximate reality, and thus a unit might see reduced SMR based simply on the resulting changes in cohort. This effect is potentially visible in the results presented comparing before and after recalibration over 2-month intervals.

The lead time bias conundrum was described in patient cohorts. When the decision to admit them to an ICU was made *before* they responded to treatment [11], mortality was underestimated. However, in resource-constrained ICUs patients are frequently selected *after* they have failed to respond to treatment. This behaviour might be better described as *treatment-failure bias*.

Thus, increasing ICU capacity relative to its demands and case mix might favourably bias the SMR over time. More generally, do admissions policies and approaches, as well as their change over time, bias SMR? Are some units therefore disadvantaged not as a result of poor care, but because of external factors less in their control?

These are potentially mischievous questions, but ones that perhaps should be examined rigorously to ensure that we meet the first goal of getting prognostic models correctly.

References

1. Juneja D, Gopal PB, Kapoor D, Raya R, Sathyanarayanan M (2011) Profiles and outcome of patients with liver cirrhosis requiring mechanical ventilation. J Intensive Care Med. doi: 10.1177/0885066611400277
2. Lin CY, Tsai FC, Tian YC, Jenq CC, Chen YC, Fang JT, Yang CW (2007) Evaluation of outcome scoring systems for patients on extracorporeal membrane oxygenation. Ann Thorac Surg 84:1256–1262
3. Beck DH, Smith GB, Pappachan JV, Millar B (2003) External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. Intensive Care Med 29:249–256
4. Lilja HE, Leppaniemi A, Kemppainen E (2008) Utilization of intensive care unit resources in severe acute pancreatitis. JOP 9:179–184
5. Minne L, Abu-Hanna A, de Jonge E (2008) Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review. Crit Care 12:R161
6. Rowan M, Ryan T, Hegarty F, O'Hare N (2007) The use of artificial neural networks to stratify the length of stay of cardiac patients based on preoperative and initial postoperative factors. Artif Intell Med 40:211–221
7. Minne L, Ludikhuize J, de Jonge E, de Rooij S, Abu-Hanna A (2011) Effect of changes over time in the performance of a customized SAPS-II model on quality of care assessment. Intensive Care Med. doi:10.1007/s00134-011-2390-2
8. Glance LG, Osler TM, Dick A (2002) Rating the quality of intensive care units: is it a function of the intensive care unit scoring system? Crit Care Med 30:1976–1982
9. Timsit JF, Fosse JP, Troche 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

10. Bakhshi-Raiez F, Peek N, Bosman RJ, de Jonge E, de Keizer NF (2007) The impact of different prognostic models and their customization on institutional comparison of intensive care units. Crit Care Med 35:2553–2560

11. Tunnell RD, Millar BW, Smith GB (1998) The effect of lead time bias on severity of illness scoring, mortality prediction and standardised mortality ratio in intensive care–a pilot study. Anaesthesia 53:1045–1053

12. Garrouste-Orgeas M, Montuclard L, Timsit JF, Reignier J, Desmettre T, Karoubi P, Moreau D, Montesino L, Duguet A, Boussat S, Ede C, Monseau Y, Paule T, Misset B, Carlet J (2005) Predictors of intensive care unit refusal in French intensive care units: a multiple-center study. Crit Care Med 33:750–755