Intensive Care Med
DOI 10.1007/s00134-017-4825-x

WHAT'S NEW IN INTENSIVE CARE

Intensive care medicine in 2050: statistical tools for development of prognostic models (why clinicians should not be ignored)

Daniele Poole[1*], Greta Carrara[2] and Guido Bertolini[2]

Predictive ability assessment of prognostic models

When physicians admit a patient to the intensive care unit (ICU), they automatically grade the degree of severity and formulate an initial prognosis. This is a complex integration process of anamnestic information and physiological data with data from experience and culture. This approach, however, does not numerically quantify the patient's risk and may be heavily influenced by subjectivity. Prognostic models, such as the severity scores of the APACHE and SAPS series, have been developed to answer this need more objectively. However, they have turned out to provide unreliable predictions when applied to contexts different from those from which they were developed [1, 2].

Prognostic models are developed on large cohorts of patients, representative of the population to which they will be applied. The prognostic weight of single clinical features (e.g. age, weight, presence of comorbidities and acute conditions on admission, physiologic parameters derangement) is measured in the development cohort with appropriate statistical procedures.

After having developed a model, its internal validity should be tested by measuring discrimination and calibration. In mortality models, discrimination is the ability to distinguish between survivors and non-survivors, while calibration measures the degree of correspondence between observed deaths and those predicted by the model across the whole range of prognostic estimates. Calibration can be visually appreciated with the GiViTI (Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva) calibration belt (Fig. 1), which provides statistically rigorous information on deviation from the ideal perfect matching between predicted and observed death rates [3, 4].

Traditionally, calibration has been assessed with the Hosmer–Lemeshow statistics [5], which, although detecting overall miscalibration, does not indicate whether the model predicts more or less deaths than those observed. The combination with traditional calibration plots does not overcome this drawback. Indeed, not being formal statistical tools, these plots provide only very rough and potentially misleading information. The GiViTI calibration belt, instead, shows exactly in which direction and for which range of risk the model miscalibrates, as illustrated in Fig. 1 and explained below. The GiViTI calibration test integrates the calibration belt, providing a p value. It is, hence, a statistical test for overall calibration, as is the Hosmer–Lemeshow statistics, which, however, corresponds exactly to the calibration belts level of statistical significance [6]. Furthermore, in two extensive simulation analyses [4, 6], the GiViTI calibration test proved to perform well and, in many situations, better than the Hosmer–Lemeshow C test, which tends to reject the null hypothesis of good calibration more often than it should.

The GiViTI calibration belt and test are available as the *givitiR* software package for R [7].

Calibration in subgroups (uniformity of fit)

Good overall calibration, however, does not grant robust calibration in important subgroups, sometimes referred to as the uniformity of fit [8].

Unfortunately, a lack of uniformity of fit is usually disregarded and unrecognised in models building a strategy.

*Correspondence: daniele.poole@alice.it
[1] Anesthesia and Intensive Care Operative Unit, S. Martino Hospital, Belluno, Italy
Full author information is available at the end of the article

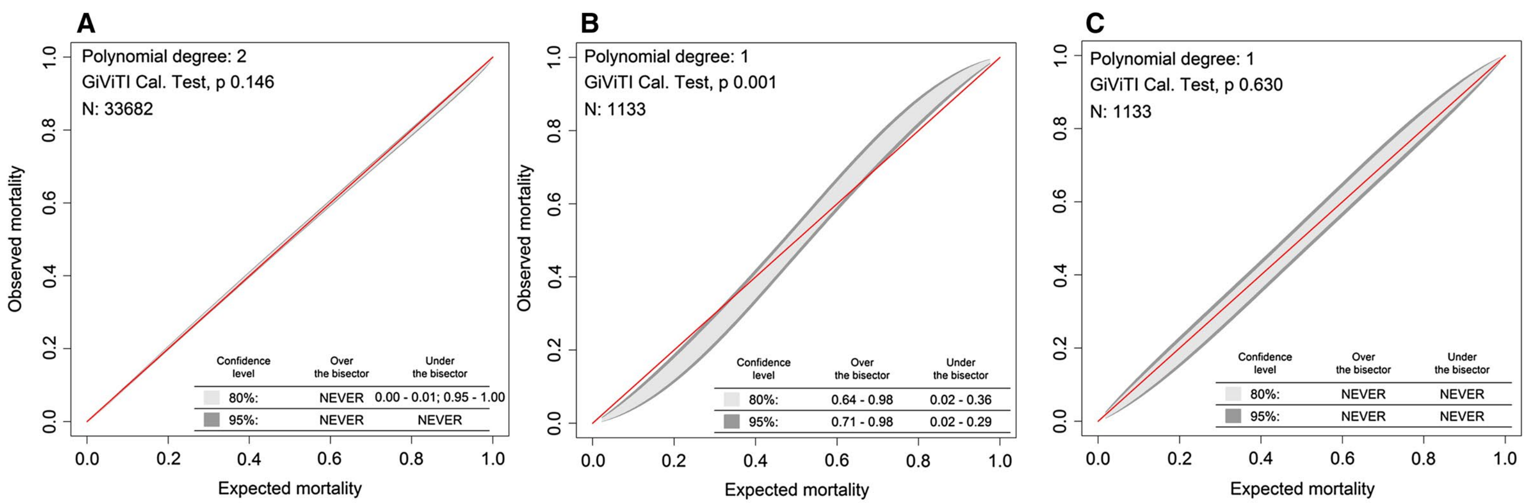

Fig. 1 GiViTI calibration belt testing a prognostic model in the overall sample, before the interaction between intra-parenchymal and GCS 3–4 was included (**a**), and in the intra-parenchymal hemorrhage subset before (**b**) and after (**c**) the inclusion of the interaction. The bisector, the *red line*, indicates the perfect correspondence between observed and predicted deaths. The *dark grey border* indicates the 95% confidence interval band and the *light grey area* the 80% confidence interval band. When the calibration belt does not include the bisector, a statistically significant miscalibration is present. In (**a**), the 95% confidence interval band includes the entire bisector indicating perfect model calibration across risk (from 0 to 100%). The band is narrow because of the high number of patients included in the analysis, providing high power for miscalibration detection. The model overpredicts in low-risk and underpredicts in high-risk patients before the inclusion of the interaction term (**b**). Calibration strikingly improves after the inclusion of the interaction is included in the model (**c**)

This is highly deplorable, as miscalibration in subgroups can seriously affect the reliability of the model when used for mortality prediction.

In Fig. 1a, we show the GiViTI calibration belt of the prognostic model developed on a cohort of 33,682 patients admitted to 178 Italian ICUs in 2012 [9]. The model calibrates very well on the overall population, as denoted by the GiViTI calibration belt always encompassing the bisector, which is the line where expected (x-axis) and observed (y-axis) mortalities match perfectly. The belt shows that, for each expected mortality rate, we have a range of possible observed rates with a 95% confidence interval that is dependent on the number of patients observed. When tested on a subset of 1133 patients with intra-parenchymal haemorrhage (a relevant prognostic factor in the model), the model miscalibrated. In the range of expected mortality rates between 2 and 29% (Fig. 1b), the band skewed downwards without including the bisector. When the band runs below this reference, it indicates that observed mortalities are significantly lower, with a 95% confidence, than those predicted by the model, indicating that the model overpredicted. The opposite happened for high-risk patients with expected mortality between 71 and 98%. In this range, the observed mortality was significantly higher than predicted, thus the model underpredicted.

As a result, in an ICU admitting mainly very severe patients with intra-parenchymal haemorrhage, the model could indicate an overall significant adjusted mortality excess (notwithstanding the good overall calibration in the development cohort), and not only in the "miscalibrating" subset. Actually, if sufficiently large, the higher prevalence of this subset in the unit (compared to the prevalence observed in the model development cohort) will magnify the effects of the "miscalibrating" subset. In this case, a poorer performance would be unfairly attributed to the ICU.

Conversely, an overprediction of overall deaths would occur if the ICU admitted mainly low-risk-of-death patients with intra-parenchymal haemorrhage. These patients would be regarded by the model as more severe than they really are, with the consequent inflation of the number of expected deaths. In this case, whatever its absolute performance, the ICU will always be unfairly rewarded in terms of a reduction of the observed to expected mortality ratio.

Poor calibration in a specific subset is thus a serious drawback of prognostic models and should be investigated in detail during the model development process. Once a lack of uniformity of fit is detected, corrective interventions should be performed. Usually, statisticians test multiplicative effects that may exist between the variables included in the model, which they call interactions. However, to selectively individuate important interactions, the contribution of clinicians is essential.

Why clinicians are so important in prognostic models development

Returning to our example, clinical experience suggests that, when intra-cerebral haemorrhage causes deep coma, the prognosis is worst. This synergism is the clinical equivalent of statistical interaction. The inclusion in the prognostic model of the interaction between intra-parenchymal haemorrhage and the Glasgow Coma Scale (GCS) 3–4, strikingly improved calibration, as shown in Fig. 1c. Now the model can be safely applied to ICUs admitting many patients with intra-parenchymal haemorrhage bearing different levels of severity. Obviously, the GiViTI calibration belt besides diagnosing a lack of uniformity of fit can be used to monitor the effectiveness of corrective interventions (Fig. 1c) [6].

There are several other clinical conditions that may have positive or negative prognostic synergisms. For example, clinicians know very well that coma has different prognostic relevance when it is the consequence of cerebral injuries or of metabolic causes. Thus, coma in chronic obstructive pulmonary diseases (COPD), for example, is not as prognostically relevant as in head trauma. Another example is septic shock due to uro-sepsis compared to other causes. In these cases, the insertion of ureteral drainage associated with antibiotic therapy is in many cases sufficient to strikingly improve clinical conditions.

The dependence of calibration from case-mix variations has been demonstrated by simulation studies and considered a good reason to develop new severity scores [10, 11]. We think, instead, that a different methodological approach should be used in prognostic models development to make them more independent of case-mix variations. This implies that the assessment of calibration in important subgroups (at least those defined by the variables included in the model) should be systematically carried out during the model development phase in order to spot and account for important clinical synergisms. Instead, researchers who have developed prognostic models overlooked the problem of biased predictions in important subgroups, limiting quality assessment to overall results. As a matter of fact, severity scores never include interaction terms, hindering the models' ability to account for case-mix differences.

Prognostic model development is a good example of how integration between statisticians and clinicians is fundamental to achieve successful results. In the GiViTI experience, close collaboration with clinicians has stimulated statisticians to develop the calibration belt and to use it to individuate clinically meaningful solutions for the improvement of model robustness. This accessible visual representation of calibration combined with formal statistical testing is a powerful tool for prognostic model development, which will be essential for future research in this field.

Author details
[1] Anesthesia and Intensive Care Operative Unit, S. Martino Hospital, Belluno, Italy. [2] GiViTI Coordinating Center, IRCCS-Istituto di Ricerche Farmacologiche 'Mario Negri', Ranica, Bergamo, Italy.

Compliance with ethical standards

Conflicts of interest
The authors declare they have no conflict of interest.

Received: 23 March 2017 Accepted: 27 April 2017
Published online: 02 June 2017

References

1. Poole D, Rossi C, Anghileri A, Giardino M, Latronico N, Radrizzani D, Langer M, Bertolini G (2009) External validation of the Simplified Acute Physiology Score (SAPS) 3 in a cohort of 28,357 patients from 147 Italian intensive care units. Intensive Care Med 35:1916–1924
2. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K (2006) Recalibration of risk prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med 34:1378–1388
3. Finazzi S, Poole D, Luciani D, Cogo PE, Bertolini G (2011) Calibration belt for quality-of-care assessment based on dichotomous outcomes. PLoS ONE 6:e16110. doi:10.1371/journal.pone.0016110
4. Nattino G, Finazzi S, Bertolini G (2014) A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. Stat Med 33:2390–2407. doi:10.1002/sim.6100
5. Lemeshow S, Hosmer DW Jr (1982) A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 115:92–106
6. Nattino G, Finazzi S, Bertolini G (2016) A new test and graphical tool to assess the goodness of fit of logistic regression models. Stat Med 35:709–720. doi:10.1002/sim.6744
7. R Core Team (2015). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/
8. Moreno R, Apolone G, Miranda DR (1998) Evaluation of the uniformity of fit of general outcome prediction models. Intensive Care Med 24:40–47
9. Rossi C, Nattino G, Facchinetti S, Fleming J, Nattino G, Nava L, Poole D, Tavola M, Bertolini G (2013) Margherita PROSAFE Project. PROmoting patient SAFEty and quality improvement in critical care. National report for general ICUs ITALY. Sestante Edizioni, Bergamo
10. Glance LG, Osler TM, Papadakos P (2000) Effect of mortality rate on the performance of the Acute Physiology and Chronic Health Evaluation II: a simulation study. Crit Care Med 28:3424–3428
11. Capuzzo M, Moreno RP, Le Gall JR (2008) Outcome prediction in critical care: the Simplified Acute Physiology Score models. Curr Opin Crit Care 14:485–490. doi:10.1097/MCC.0b013e32830864d7