Romain Pirracchio
Otavio T. Ranzani

Recalibrating our prediction models in the ICU: time to move from the abacus to the computer

Received: 15 January 2014
Accepted: 28 January 2014
Published online: 14 February 2014
© Springer-Verlag Berlin Heidelberg and ESICM 2014

R. Pirracchio (✉)
Department of Anesthesiology and Critical Care Medicine, Hôpital
Européen Georges Pompidou, Paris, France
e-mail: romainpirracchio@yahoo.fr
Tel.: +33-1-56092513

R. Pirracchio
Division of Biostatistics, Hôpital Saint Louis, Paris, France

O. T. Ranzani
Hospital das Clínicas, University of São Paulo, Rua Dr. Enéas
de Carvalho Aguiar, 255, São Paulo 05403-900, Brazil

O. T. Ranzani
Amil Critical Care Group, Intensive Care Department, Hospital
Paulistano, Rua Martiniano de Carvalho, 741,
São Paulo 01321-001, Brazil

Delirium is well recognized as a risk factor of adverse outcomes in critically ill patients. However, the best strategy for prevention and treatment is still debated. Early identification of the patients at high risk of developing a delirium could help to prevent the symptom.

The PRE-DELIRIC score was proposed by van den Boogaard et al. [1] in 2012. This score exhibited good prediction performances, but was trained on a sample encompassing only five ICUs in the Netherlands, and thus might lack external validity. Validation of prediction models in different study populations is necessary to increase generalizability, because a prediction model can show myriad patterns of invalidity when tested in new

settings and also have its performance decline over time. Therefore, in this issue of *Intensive Care Medicine*, the same group reports the results of a study that aimed at recalibrating the PRE-DELIRIC model using a large multinational data set [2]. Their main result is that the distribution of the predictors included in the score highly differed between centers and thus recalibration resulted in better prediction performances.

When appraising the performances of a prediction model, the two crucial dimensions are discrimination and calibration. Discrimination is a measure of how well the model is able to distinguish between patients who will or will not present the outcome. The area under the receiver operating characteristic curve (AUROC or *c*-statistic) is the most widely used measure for model discrimination. Calibration is a measure of the agreement between the predicted and the observed outcomes. A well-calibrated model implies that, on average, the predicted risk within subgroups matches the proportion of subjects that actually present the outcome in the subgroup. To assess model calibration, the most popular method is the Hosmer–Lemeshow statistic. It is also possible to plot the estimated risk (*x* axis) vs. the observed outcomes (*y* axis). A perfect calibration is an identity line (45 °) and, in general, the distance between the prediction and the identity lines is representative of the model calibration. Calibration plots can be analyzed using different perspectives. Calibration-in-the-large (e.g., mean calibration) assesses whether prediction agrees on average with observed probabilities. We also can fit a linear regression between both probabilities, retrieving a calibration slope. The results from the calibration-in-the-large (mean difference/intercept—parameter *a*) and the calibration slope (parameter *b*) offer a good evaluation of the model calibration. An accurate evaluation of the calibration properties is crucial as model miscalibration is likely to cause systematic errors in the decision-making process. This is why model recalibration is a priority topic. When

Table 1 Common measures for calibration and recalibration approaches of prediction models (adapted from Steyerberg [12])

Calibration measure ^a	Aim	Values/output	Advantages	Disadvantages
Hosmer–Lemeshow test	Comparing observed vs. predicted event rates, grouping patients by ranges of expected risk (usually 10 subgroups)	C and H statistics and p values; if significant, the observed and predicted probabilities are different; the model is not well calibrated	Conceptually easy Basic computation	Limited power Highly dependent on the choices of how predictions are grouped
Calibration-in-the-large	Compare the average observed event (Y) vs. the average predicted event (Y')	Based on a calibration plot, where the value of interest is the intercept (a) of a linear regression	Very useful for model validation: poor calibration is associated with systematic errors Useful to detect overfitting	Does not explore overfitting Comparison of average predicted vs. observed risks. Some models which are well “calibrated-in-the-large” could be not useful for certain ranges of predictions
Calibration slope	Estimate of the mean strength of association between the average observed event (Y) vs. the average predicted event (Y')	Based on a calibration plot, where the value of interest is the slope (b) of a linear regression	Key measure in the validation setting, indicating how much, on average, we need adjust the effects of predictors to make the model calibrated	Represents the average strength between predictors and outcomes. Might poorly explore some specific ranges of prediction
Calibration belt	Creates a confidence band for the calibration curve based on a function that relates expected to observed probabilities across classes of risk	Through a calibration plot, fits the predicted data with a general calibration function, and plots the corresponding curve from the function	Does not rely on risk ranges (unlike Hosmer–Lemeshow test) Provides an estimate of the degree of uncertainty around the calibration curve Does not use average risks by grouping patients, providing more information about individual contribution to the model	Needs further validation in new data, mainly regarding the influence of the sample size and the theoretical limits from the function Requires deeper mathematics
Harrell’s E statistic	Calculates the absolute difference within smoothed Y' and the identity line (ideal calibration)	Through a calibration plot, fits the predicted data using a smoothing function	Conceptually easy Summarizes miscalibration over the whole curve Provides an estimate of the degree of uncertainty around smoothed line and identity	Dependence on the smooth algorithm
Computation				
Recalibration approach ^b Intercept re-estimation	To correct the calibration-in-the-large, i.e., adjust the average difference between observed vs. predicted outcomes in the new setting		Fixing the slope b of 1, re-estimates the intercept	
Slope re-estimation	To correct the calibration slope, i.e., decrease the overfitting		Fixing the intercept, re-estimates the value of a new slope b . Then, applies the new b to the original model, thereby using shrunk coefficients	
Updating approach Model revision	After recalibration, revises the effect of each predictor in the new data		Refits the original model, first evaluating the effect of strongest predictors in the new setting and adjusting in accordance	
Model extension	After recalibration, builds a new hybrid model, aggregating previous published associations from the literature		Refits the original model, testing if new predictors are able to improve the model performance	

^a This table shows some measures of calibration (several others exist) and depicts a summary of each one^b Regarding recalibration/updating models, all the approaches could be used simultaneously

recalibrating a prediction model, one can re-estimate the intercept (parameter a), the slope (parameter b), or both in different populations. The original model can also be refitted and some predictors might be included/excluded before to re-estimate model calibration. It is also possible to build a new “hybrid” model, aggregating previous published associations [3]. Details about calibration and recalibration procedures are provided in Table 1.

In their updated study, van den Boogaard et al. chose to recalibrate the PRE-DELIRIC model without revising the original model [2]. They used four approaches to recalibrate the model, working on the concept of the slope and the intercept in a calibration plot. The authors proposed to multiply each coefficient in the first model by the new slope and used the new intercept. The recalibrated PRE-DELIRIC model showed a similar discrimination (AUROC) and an improved calibration. It should be noted that the former PRE-DELIRIC model already had an acceptable performance in the multinational sample, as a result of a very well conducted score development in the first study [1]. However, the recalibration study [2] suffers some weaknesses. The most important is the lack of external/internal validation procedure. Internal validity procedures aim at improving the prediction performances, together with limiting the risk of overfitting [3]. Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Several internal validity procedures are available, such as bootstrapping techniques or cross-validation [3]. Whereas in the seminal study the authors used bootstrap [1], no internal validity procedure was used in the recalibration study [2]. Because of the lack of internal validation, the reported performances for the recalibrated model are likely to be overestimated. External validity procedures aim at guaranteeing that the prediction performances can be generalized to other populations. The lack of diversity of the population used to train the model is a serious threat to external validity. In the present case, the authors used multinational data to recalibrate the model and thus to increase its generalizability. However, an adequate evaluation of the prediction performances would have been based on an external data set [3].

In its former and recalibrated version [1, 2], the PRE-DELIRIC score relies on a logistic regression. Although generalized linear models are widely used for prediction in medicine, such models impose stringent constraints on the relationship between explanatory variables and the outcome. Given the complexity of the processes underlying disease development in ICU patients, this assumption might be unrealistic. If this assumption fails to hold, all derived estimates of the risk of death are necessarily biased and predictive power may be relatively low. Hence, limited calibration of the PRE-DELIRIC

score might be to some extent related to the misspecification of the underlying statistical model. The fact that mixed effect models, which add some flexibility and randomness to the modeling process, clearly outperformed standard main-term logistic regression supports such a hypothesis.

We can reasonably expect that prediction can be improved by using an automated algorithm to estimate risk of delirium without requiring any specification about the shape of the underlying relationship. Such methods are often referred to as nonparametric. Some studies have already been used to fit ICU data [4–9] and unanimously concluded that nonparametric methods might perform at least as well as standard logistic regression. However, there are numbers of different nonparametric algorithms, and there is no rule to guide the choice of the appropriate method. In order to avoid unjustified choices, we would like to let the data decide which algorithm to use. Recently, the super learner was developed as a data-adaptive technique for selecting an optimal regression algorithm among a set of candidate algorithms [10, 11]. To implement the super learner, a user must provide a customized collection of algorithms. The super learner estimates the risk associated with each algorithm using cross-validation, from which it builds an aggregate algorithm obtained as the optimal weighted combination of the candidate algorithms. The performance of any given algorithm will generally vary widely among various settings. Theoretical results have nonetheless demonstrated that the super learner performs no worse than the optimal choice among the provided library of candidate algorithms [11]. This is significant because in a given application the true optimal choice cannot be known. The super learner capitalizes on the richness of the library it builds upon and generally offers gains over any specific candidate algorithm in terms of flexibility to accurately fit the data.

There is a potential interest in regularly recalibrating previously published prediction scores, especially when the former models were trained and validated locally. However, despite repeated recalibrations, it is very likely that the prediction performances will reach a steady state, where no further gain can be expected from further recalibrations based on parametric models. Moving from such stringent modeling approaches to flexible data-adaptive algorithms is probably the way to exceed such limits. Things that were unthinkable at the time of the abacus are now possible. Let us dive into the future of modeling!

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

References

1. Van den Boogaard M, Pickkers P, Slooter AJC et al (2012) Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICU patients) delirium prediction model for intensive care patients: observational multicentre study. *BMJ* 344:e420
2. Van den Boogaard M, Schoonhoven L, Maseda E et al (2014) Recalibration of the delirium prediction model for ICU patients (PRE-DELIRIC): a multinational observational study. *Intensive Care Med*. doi: [10.1007/s00134-013-3202-7](https://doi.org/10.1007/s00134-013-3202-7)
3. Steyerberg EW, Harrell FE Jr, Borsboom GJJ et al (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 54:774–781. doi: [10.1016/S0895-4356\(01\)00341-9](https://doi.org/10.1016/S0895-4356(01)00341-9)
4. Dybowski R, Weller P, Chang R, Gant V (1996) Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *Lancet* 347:1146–1150
5. Clermont G, Angus DC, DiRusso SM et al (2001) Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. *Crit Care Med* 29:291–296
6. Ribas VJ, López JC, Ruiz-Sanmartín A et al (2011) Severe sepsis mortality prediction with relevance vector machines. *Conf Proc IEEE Eng Med Biol Soc* 2011:100–103. doi: [10.1109/IEMBS.2011.6089906](https://doi.org/10.1109/IEMBS.2011.6089906)
7. Kim S, Kim W, Park RW (2011) A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthc Inform Res* 17:232–243. doi: [10.4258/hir.2011.17.4.232](https://doi.org/10.4258/hir.2011.17.4.232)
8. Foltran F, Berchialla P, Giunta F et al (2010) Using VLAD scores to have a look insight ICU performance: towards a modelling of the errors. *J Eval Clin Pract* 16:968–975. doi: [10.1111/j.1365-2753.2009.01240.x](https://doi.org/10.1111/j.1365-2753.2009.01240.x)
9. Gortzis LG, Sakellariopoulos F, Ilias I et al (2008) Predicting ICU survival: a meta-level approach. *BMC Health Serv Res* 8:157. doi: [10.1186/1472-6963-8-157](https://doi.org/10.1186/1472-6963-8-157)
10. Dudoit S, Van Der Laan MJ (2003) Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat Methodol* 2:131–154
11. Van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. *Stat Appl Genet Mol Biol* 6:Article 25. doi: [10.2202/1544-6115.1309](https://doi.org/10.2202/1544-6115.1309)
12. Steyerberg EW (2009) *Clinical prediction models: a practical approach to development, validation, and updating*. Springer, New York