Mortality risk prediction models: Methods of assessing discrimination and calibration and what they mean

L J Solomon, MB ChB, MMed, Cert (Paed), Critical Care (SA)

Department of Paediatrics and Child Health, University of the Free State, Bloemfontein, South Africa

Corresponding author: *L J Solomon (SolomonLJ@ufs.ac.za)*

Mortality prediction models, which this author prefers to call outcome-risk assignment models, are ubiquitous in critical care practice. They are designed to assess the risk of dying for any given patient by assigning risk based on indices of physiological derangement, high-risk diagnoses or diagnostic categories, or the need for therapeutic intervention to support organ function. It is important to recognise that predictive models derived to date are not accurate enough to allow reliable prediction of individual patient outcome, and therefore cannot be used as criteria for admission to intensive care. Patients identified as high risk for mortality must be flagged for appropriate intensive intervention and close monitoring in order to reduce the risk of dying due to the current illness.

South Afr J Crit Care 2022:38(1):2-3. https://doi.org/10.7196/SAJCC.2022.v38i1.548

Mortality-risk prediction models are used to benchmark ICU performance against 'gold standards', being the ICUs where the models had been derived at the time they were derived. Therefore, it is important to apply up-to-date risk prediction models.[1] ICU performance metrics based on mortality risk prediction models include effectiveness, measured by the standardised mortality ratio (SMR) and efficiency, measured by the number of patients requiring at least one ICU-care modality and mortality risk >1%.[2] They are also used as the basis for risk-adjusted control chart methodologies which track ICU performance over time[3] and for quality improvement efforts, for example, where the presence of intensivists has reduced SMR[4] and where excessive deaths among low-risk patients were related to invasive procedures and inadequate infection control practices.[5]

Model derivation methods have evolved over time. Most models are derived by a combination of expert selection of variables subjected to univariate and multivariate logistic regression analysis. This process produces a list of independent variables with their related coefficients and odds ratios for the dependent variable, or outcome, usually death or survival. Generally, these models generate a score from the logistic function which is in turn transformed to a mortality risk at individual patient level. Therefore, in a population of patients, the sum of individual mortality risks generates the expected number of deaths.

Neural network-based machine learning models are emerging and show promise to perform as well as, if not better, than statistical models.[6] Better performance, however, may come at the not insurmountable cost of added complexity and the need for access to appropriate computational resources.[7]

However they are derived, models need to perform well in both the derivation cohort and in the intended use-case context. This is important to establish prior to investing the time, effort and money it takes to deploy these models in any ICU. The models need to demonstrate good discrimination, defined as the ability to appropriately classify all patients such that the observed and predicted outcome rates are as close as possible. Receiver Operator Characteristic (ROC) curve analysis is commonly used to assess model classification ability. Essentially, ROC curves plot truly predicted non-survivor rates against the falsely predicted non-survivor rates for each value of the score. The ideal classifier, which does not exist, would have an area under this curve (AUC) of 1. A poor (random) classifier would have an AUC of 0.5. Classifiers are regarded as acceptable for AUCs from 0.7 to <0.8, good for AUCs of 0.8 to <0.9 and excellent if ≥ 0.9.[1]

ROC analysis applied to datasets with unbalanced dependent outcomes, such as mortality and survival, however, might show overly optimistic AUC values because a hypothetical model, for example, which categorises all cases as survivors in a population with a 10% mortality rate, would have an AUC of 0.9.

An alternative strategy to assess model discrimination on imbalanced datasets is to consider model precision as positive predictive value, the ratio of true positive class and model *predicted* positive class, and recall as sensitivity or true positive rate, the ratio of true positive class to *actual* or observed positive class. The associated precision/recall curve (PRC) relates precision to recall.[8] Similar to ROC curves, the PRC plot generates a curve with an AUC. The greater the AUC, the better the discrimination. Again, perfect discrimination would yield an AUC of 1. Random model performance determined by PRC depends on the degree of class imbalance. This 'baseline' on the y-axis of the plot (horizontal to the x-axis) is calculated as the ratio of positive class to the sum of positive and negative classes (y=P/(P+N)). AUC of PRC curves which indicate random model performance will therefore also vary with class balance and will be equal to y.

Other measures of model performance include accuracy, defined as the number of true positive and true negative classes within the whole sample (TP+TN)/(TP+FP+TN+FN) and the F1 Score which is the harmonic mean of precision and recall: 2*(precision*recall)/(precision + recall).[9]

Models also need to perform well within categories of mortality risk, different demographics, and across diagnostic groups i.e., model calibration. The Hosmer-Lemeshow (H-L) statistical analysis is widely used to assess model calibration, by determining the significance (*p*-value) of the differences between observed and expected outcome rates within, usually, 10 groups (deciles) of increasing mortality risk. This methodology is valid only if a few provisos are met, facilitated by sufficiently large data sets.[1] H-L would be unreliable if sample size is less than 400 or if >4 out of 20 values of the 'expected' columns of the H-L table are <5. The same *p*-value may be found if there is a small difference in a large sample, as with a large

difference in a small sample. Therefore, the *p*-value per se says nothing about the clinical significance of a difference between observed and expected mortality rates. Careful inspection of the H-L table will yield a better 'idea' of the nature of the lack-of-fit. The H-L test will almost always show a significant lack-of-fit if the observed vs expected mortality rates are not similar in all deciles of risk. 95% confidence intervals (CI) for SMR must be narrow for it to have meaning. This depends on the number of recorded deaths. CIs will be wide if <50 are recorded.

Shann[1] states that when interpreting lack of calibration, it is more likely to mean that the standard of care in the unit is different than the units in which the model was derived, at the time that it was derived. Thus, if SMR is >1, care is worse. If SMR is <1, care is better. Based on ROC and H-L analysis, Shann[1] further suggests a guide for deciding whether a model is appropriate for any context. If the ROC AUC is >0.7 and similar numbers or proportions of observed v. expected outcomes are found across all H-L deciles of risk, the model is appropriate. Conversely a model may not be appropriate if AUC-ROC is <0.7 or if more deaths occur than predicted in lower risk categories and less deaths than predicted occur in higher risk categories.[1] However issues of case mix and resource differences between derivation and application contexts need to be considered.[10]

Owing to the limitations of the H-L method, which include artificial grouping of data into risk strata, a *p*-value which is indifferent to the type or extent of miscalibration and that H-L suffers from low statistical power,[6,11] alternative tests of discrimination need to be considered. The flexible calibration curve with its slope and intercept is considered a superior, though less popular, assessment of model calibration.[11] Model calibration thus assessed can be classified either as mean, weak, moderate or strong. Mean calibration refers to the ability of the model to predict on average the outcome of interest. Over-prediction occurs when the mean predicted outcome is greater than observed, while under-prediction occurs when the average predicted outcome is less than observed.

Weak calibration is defined by a flexible curve with a slope that is either greater or less than 1. If the slope is less than 1 then the model over-predicts, or under-predicts if >1. The intercept of the flexible curve indicates over-prediction or under-prediction at values less than or greater than zero, respectively.

Moderate calibration is assessed by comparing how well the calibration curve fits the model prediction to the observed outcomes. The flexible curve will be close to the diagonal when proportions of predicted and observed outcomes are similar.

In the current *SAJCC* issue, Pazi *et al.*,[12] using a cohort of 829 patients admitted to an adult ICU in a tertiary hospital in South Africa (SA), have developed a SAPS III-based mortality prediction model calibrated to their specific ICU population using data available from a previous SAPS III validation study.[13] Data were collected for one year starting in January 2017. They reported a mortality rate of 21.35%.

The rationale for this unit-specific model derivation was the age of SAPS III[14] and the fact that no data from centres in low- to-middle income countries were included in the derivation dataset, and none from SA. They employed logistic regression to select variables for four models which where than subjected to cross-validation analyses and a final model was internally validated using measures of discrimination (ROC analysis, precision recall, balanced accuracy, bookmarked informedness and markedness) and calibration (H-L and the flexible calibration curve).

The authors report a ROC-AUC of 0.86 and PRC-AUC of 0.67 with a baseline of 0.2. Therefore global model discrimination is deemed good. Calibration as determined by H-L, *C* and *H*-statistics produced *p*-values of 0.95 and 0.93, respectively. However, SMRs are not consistent among all risk categories and the associated CIs are consistently wide, including unity. This illustrates the concerns expressed with regard to the H-L being a reliable measure of model calibration. The authors further present the flexible calibration curve with its slope and intercept and 95% CIs. The intercept, though negative, is close to zero with 95% CIs -0.44 - 0.27. Similarly, the slope of the curve is 1.04 with a 95% CI 0.76 - 1.36. The model therefore shows moderate calibration, using the above criteria, with a tendency to 'oscillate' within the CI limits.

External validation of this model is needed, as pointed out by the authors, preferably on larger data-sets among a range of ICUs in SA, before it can be accepted as a national standard for benchmarking performance of adult ICUs.

In conclusion, understanding model validation methods will promote appropriate mortality risk assignment model derivation, choice, validation and application which in turn could increase the confidence clinicians have in the ICU performance metrics that these models facilitate, and ultimately that systems of care can be designed which maximise best outcomes for patients in ICUs where these models are deployed.

1. Shann F. Are we doing a good job: PRISM, PIM and all that. Intensive Care Med 2002;28(2):105-107. https://doi.org/10.1007/s00134-001-1186-1
2. Gemke RJ, Bonsel GJ, van Vught AJ. Effectiveness and efficiency of a Dutch pediatric intensive care unit: validity and application of the Pediatric Risk of Mortality score. Crit Care Med 1994;22(9):1477-1484. https://doi.org/10.1097/00003246-199409000-00020
3. Baghurst PA, Norton L, Slater A, ANZICS Paediatric Study Group. The application of risk-adjusted control charts using the Paediatric Index of Mortality 2 for monitoring paediatric intensive care performance in Australia and New Zealand. Intensive Care Med 2008;34(7):1281-1218. https://doi.org/10.1007/s00134-008-1081-0
4. Goh AY, Lum LC, Abdel-Latif ME. Impact of 24 hour critical care physician staffing on case-mix adjusted mortality in paediatric intensive care. Lancet 2001;357(9254):445-446. https://doi.org/10.1016/S0140-6736(00)04014-9.
5. Earle M, Natera OM, Zaslavsky A, et al. Outcome of pediatric intensive care at six centers in Mexico and Ecuador. Crit Care Med 1997;25(9):1462-1427. https://doi.org/10.1097/00003246-199709000-00011
6. Keuning BE, Kaufmann T, Wiersema R, et al. Mortality prediction models in the adult critically ill: A scoping review. Acta Anaesthesiol Scand 2020;64(4):424-442. https://doi.org/10.1111/aas.13527
7. Pienaar MA, Sempa JB, Luwes N, Solomon LJ. An artificial neural network model for pediatric mortality prediction in two tertiary pediatric intensive care units in South Africa. A development study. Front Pediatr 2022;10:797080. https://doi.org/10.3389/fped.2022.797080
8. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432
9. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006:1015-1021.
10. Solomon LJ, Naidoo KD, Appel I, et al. Pediatric index of mortality 3—an evaluation of function among ICUs in South Africa. Pediatr Crit Care Med 2021;Publish Ahead of Print. https://doi.org/10.1097/pcc.0000000000002693
11. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019;17(1):230. https://doi.org/10.1186/s12916-019-1466-7
12. Pazi S, Sharp G, van der Merwe E. Prediction of in-hospital mortality: an adaptive severity of illness score for a tertiary ICU in South Africa. South Afr J Crit Care 2022;38(1):4-9.
13. van der Merwe E, Kapp J, Pazi S, et al. The SAPS 3 score as a predictor of hospital mortality in a South African tertiary intensive care unit: A prospective cohort study. PLoS One 2020;15(5):e0233317. https://doi.org/10.1371/journal.pone.0233317
14. Moreno RP, Metnitz PGH, Almeida E, et al. SAPS 3 - from evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med 2005;31(10):1345-1355. https://doi.org/10.1007/s00134-005-2763-5

Accepted 24 March 2022.