How to externally validate prognostic models in surgery

DW de Lange

Department of Intensive Care, University Medical Center Utrecht, The Netherlands

Keywords - Prediction mortality models, ICU prediction mortality model validation, ICU outcome, outcome prediction Intensive care, long-term survival

In this issue of *Netherlands Journal of Critical Care*, Timmers et al. [1] show that applying severity of illness scores (like APACHE, SAPS or MPM) to a population other than originally intended is difficult and has certain limitations. So, if this is the case, why do we have such prognostic models and why do we want to apply them to different subgroups? As early as 1981, with the introduction of the APACHE model, it became clear that prognostic models are not designed to assist physicians in making decisions on individual patients, but rather to classify groups of patients on the basis of their severity of illness [2,3]. We use those severity scores to compare groups of patients and detect trends within such groups or across institutions. This is called benchmarking.

Benchmarking is a term borrowed from industry and means "searching for industrial best practices that lead to superior performance" [4]. In the 1960s the Xerox Company developed a popular copying machine. The boss of Xerox, Charles Christ, saw an advertisement promoting a cheaper model. He wanted "a benchmark, something I can measure myself against to understand where we have to go from here". This led to the formalization of a benchmarking programme [5]. Hence, the important aim of benchmarking is to learn and improve through comparison with others. In critical care, benchmarking has been used to compare outcome and thus continuously improve quality. For this purpose of comparing outcome, several severity of illness prognostic models have been developed.

The most widely used severity of illness prognostic models (APACHE II and SAPS II, MPM0 and MPM24) are all based on logistic regression analysis. All these models have one ultimate goal: to accurately predict mortality, even outside the setting in which the score has originally been developed [3].

If a prognostic model includes the most important variables and is not seriously over fitted, it should adequately predict mortality in a setting with a different case-mix. However, we do not know if all important variables have been included in the model. Therefore, even if the model has a good calibration and discrimination assessed by internal validation, it does not necessarily mean that the model predicts well in a different context (so-called external validation).

The above means that the choice of the patients in which the external validation takes place influences the performance of a prognostic model. For instance, APACHE IV was developed in the United States in a cohort of over 100,000 patients. Recent external validation of this model in the Netherlands showed a large overprediction (meaning less observed mortality than predicted mortality) across all deciles of risk (see Figure 1). Fortunately, this external validation was performed on a very large database that included 65 different ICUs [6]. In this issue, however, Timmers et al. describe an external validation of different prognostic models in a subset of surgical patients in only one hospital [1]. The calibration plots of the APACHE II show an underprediction of mortality across the lower deciles of risk (meaning more observed than predicted mortality). Because only one hospital is involved in this study, it is difficult to say whether the model has a low calibration or the hospital is a poor performer! In general, the target population should be predefined to permit selection of a representative sample at the time of the model's development. In other words, we should know the type of patients we want to apply the model to before developing it. Using it for subsets of patients might diminish calibration and discrimination [7].

Even when a model is developed on a general ICU population, like APACHE IV and applied to a general ICU population in the Netherlands, it still needs to be validated [6]. If such a prognostic model does not perfectly fit the Dutch situation, it needs to be customized (recalibrated). In general, there are two ways to customize a prognostic model. First-level customization includes the original variables of the model but gives them new weights (coefficients) that better fit the Dutch situation. This actually means that the entire model is rebuilt. Second-level customization leaves the model and the coefficients unchanged but corrects the weight of the entire model (see Figure 1). This can only be done when the original prognostic model constantly under- or overpredicts mortality across all deciles of risk [8]. In their study, Timmers et al. use intercept adjustment to customize the model to better match the observed mortality in their cohort [1]. Intercept adjustment is a form of second-level customization. However, the Timmers et al. study shows that APACHE II, SAPS II and MPM II do not constantly underpredict or overpredict mortality across all deciles of risk (see Figure 2). Not surprisingly, intercept adjustment was unable to make those models fit perfectly in their study. However, customization did make the
models fit better for the majority of the patients in this study who were in the lower deciles of risk (see Figure 3 of the manuscript) [1].

Another problem with external validation studies is sample size. Previous studies have shown that all statistical tests to ascertain discrimination (e.g. Brier-score, Hosmer-Lemeshow) are unreliable when sample size does not exceed 2500 ICU patients (i.e. >500 outcome events or deaths) [9]. Unfortunately, the sample of surgical patients in the study by Timmers et al. was only 1822 ICU patients. Although the authors suggest that APACHE II and SAPS 3 are the models that best predict outcome in this surgical subpopulation, we are not certain that the other models in their study are inferior.

When the outcome of an external validation is unsatisfactory, authors often explain this as being caused by differences in case-mix between the original population and the subpopulation used for the external validation. In the study by Timmers et al., the surgical population is clearly different from the mixed ICU-population that was used to create the APACHE II model. Differences in case-mix can lead to differences in outcome prevalences, which might corrupt the models’ predictive ability. Another explanation for a model’s declining predictive ability is age. For example, APACHE II was introduced in 1984, over 26 years ago! A model is outdated as soon as it has been developed. For instance, APACHE IV has been available since 2005, but the model was based on patients admitted to American ICUs in 2002-2003. However, this model was validated for the Dutch population in 2010 [6]. Hence, the best prognostic model available to us is already 8 years old! For benchmarking purposes it is not only the age of the model that is important. The model needs to be customized yearly to best fit the general Dutch population at that time. Only then can it be used for benchmarking purposes.

Clearly, there are some pitfalls associated with external validation of prognostic models, especially in subgroups, such as surgical ICU patients. However, waiting for the ideal model that perfectly fits Dutch surgical ICU patients will takes decades and will be outdated once it has been developed. Apparently, one of the oldest models (APACHE II) best predicted 5-year-survival in this surgical cohort from 1995-2000 [1].

### Figure 1. Terminology and Abbreviations

**APACHE:** Acute Physiology and Chronic Health Evaluation.

**SAPS:** Simplified Acute Physiology Score

**MPM:** Mortality Prediction Model

The term **discrimination** refers to a model’s ability to distinguish survivors from non-survivors. As a measure of discrimination often the area under the receiver operating curve (ROC) is calculated. This area under the curve (AUC; sometimes called C-index) represents the probability that an arbitrary patient who died had a higher predicted risk than an arbitrary patient who survived. An AUC of 0.5 indicates that the model does not predict better than chance. An AUC of 1 indicates that the model discriminates perfectly.

**Accuracy** refers to the difference between predicted risks and observed outcomes at the level of individuals. Often the Brier inaccuracy score is used, which is the mean of the squared error. If you predicted that a patient had a 10% chance of dying but he dies anyway, the difference is (1-0.1)=0.9 (poor prediction). If a patient has a 90% chance of dying and he dies, the difference is (1-0.9)=0.1 (quite well predicted). So, the Brier score is the mean of all these differences and returns a value between “0” (perfect) and “1” (no fit).

The concept of **calibration** refers to the agreement between observed and predicted risks. Because we cannot observe risks directly (we only know whether a patient died or not), calibration can only be measured indirectly. Two different approaches to measure calibration are often used. The first approach was proposed by Cox and uses logistic regression to verify the agreement between predicted and observed risks. The second approach was proposed by Hosmer and Lemeshow, and builds on the calibration statistic C. This statistic was designed to assess whether a given logistic regression model fits to a particular data set.

When a predictive model calibrates poorly on an external data set, one may try to improve its performance by customizing (re-calibrating) the model to the data. Several strategies exist for this purpose. For instance, you may choose to re-estimate a model’s coefficients on the new data, and to add or remove terms from the model. A simpler customization strategy is to re-estimate the intercept and slope of the linear predictor, by fitting a new logistic regression equation with observed outcome as the dependent variable and the logit-transformed original predictions as the independent variable; this has been termed “logistic recalibration”.

Definitions are based on the article by Peek et al. [9] In this article more examples are given and limitations of the statistical methods are discussed.
Comparing groups of patients from year to year or across institutions with the intent of improving quality of care (benchmarking) does not only rely on good external validation of modern prognostic models. A permanent, national, high-quality database is pivotal for improving the quality of ICU care. This database (NICE) is yearly updated (customized and recalibrated) and thus compensates for ageing of the prognostic models. Additionally, a national database is less influenced by unmeasured variables that are different in the Dutch ICU situation from the original population: e.g. differences in culture, admission policies, do-not-resuscitate orders, etc. In the Netherlands we are quite fortunate that the vast majority of ICUs (more than 85%) participate in the National Intensive Care Evaluation (NICE) database. Having such a large database (over 65,000 admissions yearly) enables NICE to constantly recalibrate prognostic models and provide accurate benchmarking data.

References

1. Timmers TKD, Verhoefstad MHJ, Moons KGM, Leenen LPH. Validation of six mortality prediction systems for surgical population admitted to the ICU. Neth J Crit Care 2011; 118-130.