ELSEVIER

Journal of Clinical Epidemiology 60 (2007) 491–501

Journal of Clinical Epidemiology

External validation of prognostic models for critically ill patients required substantial sample sizes

N. Peek[a,*], D.G.T. Arts[b], R.J. Bosman[c], P.H.J. van der Voort[c], N.F. de Keizer[a]

[a]*Department of Medical Informatics, Academic Medical Center – Universiteit van Amsterdam, Amsterdam, the Netherlands*

[b]*Austrian Health Institute, Vienna, Austria*

[c]*Department of Intensive Care, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands*

Accepted 23 August 2006

Abstract

Objective: To investigate the behavior of predictive performance measures that are commonly used in external validation of prognostic models for outcome at intensive care units (ICUs).

Study Design and Setting: Four prognostic models (Simplified Acute Physiology Score II, the Acute Physiology and Chronic Health Evaluation II, and the Mortality Probability Models II) were evaluated in the Dutch National Intensive Care Evaluation registry database. For each model discrimination (AUC), accuracy (Brier score), and two calibration measures were assessed on data from 41,239 ICU admissions. This validation procedure was repeated with smaller subsamples randomly drawn from the database, and the results were compared with those obtained on the entire data set.

Results: Differences in performance between the models were small. The AUC and Brier score showed large variation with small samples. Standard errors of AUC values were accurate but the power to detect differences in performance was low. Calibration tests were extremely sensitive to sample size. Direct comparison of performance, without statistical analysis, was unreliable with either measure.

Conclusion: Substantial sample sizes are required for performance assessment and model comparison in external validation. Calibration statistics and significance tests should not be used in these settings. Instead, a simple customization method to repair lack-of-fit problems is recommended.

Keywords: Prognostic models; Validation studies; Sample size; SAPS II; APACHE II; MPM II

1. Introduction

Prognostic models are important tools to provide estimates of patient outcome probabilities. Within the field of intensive care (IC) medicine, prognostic models are often used for mortality predictions which enable, for example, the stratification of patients for enrollment in clinical trials and controlling for severity of illness in auditing quality of care [1,2]. Four well-known prognostic models in IC are the Simplified Acute Physiology Score II (SAPS II) [3], the Acute Physiology and Chronic Health Evaluation II (APACHE II) [4] and the Mortality Probability Models II (MPM_0 II and MPM_{24} II) [5]. All four models are logistic regression models to predict the probability of in-hospital mortality. They use slightly different sets of covariates describing the demography (e.g., age), admission type (e.g., medical, urgent surgical) comorbidity (e.g., chronic dialysis, respiratory insufficiency), and worst physiological status of the patient in the first 24 hours of IC admission (e.g., highest body temperature, lowest blood pressure), or, in case of the MPM_0 II, in the first hour of IC admission. In the Appendix, a more extensive description of the four models is given.

In many countries, regional or national registries have been established that use one or more of these four prognostic models to audit the quality of IC medicine [6,7]. One example is the National Intensive Care Evaluation (NICE) registry that aims to assess and improve the quality of intensive care units (ICUs) in the Netherlands [8]. Because the IC prognostic models were developed 20 years ago on other (American or European) patient populations than those to which they are applied now, their generalizability must be assessed before the models can be used in clinical practice [2,9]. Therefore, many studies have been published on validating and comparing these models in external settings, with the aim of choosing the best

* Corresponding author. Tel.: 31 20 5667872; fax: 31 20 6919840.
E-mail address: n.b.peek@amc.uva.nl (N. Peek).

0895-4356/07/$ – see front matter
doi: 10.1016/j.jclinepi.2006.08.011

performing model and to assess its performance. These studies commonly focus on measuring the models' discrimination using the area under the Receiver Operating Characteristic (ROC) Curve [10], and their calibration using the Hosmer–Lemeshow goodness-of-fit statistics [11].

The results of these studies vary considerably. Whereas one study [12] concludes that the discrimination of the SAPS II model is superior to that of the APACHE II model, another [13] does not find a difference in discriminative ability between the two models. Similarly, some studies conclude that based on the measured calibration the SAPS II model is insufficient [14,15], whereas another concludes that calibration of the SAPS II is sufficient [16]. The variation in these results might be caused by temporal or geographical differences between the data sets that were used. However, the variation in results might also be caused by random variation in the validation samples. Estimates of predictive performance in external data (i.e., sampled from a different population than the data that was used to derive the model) are known to be highly variable [17]. The numbers of observations in the data sets that were used in the studies mentioned above vary widely, from 300 to 16,000.

The goal of this study was to validate and compare the performance of the APACHE II, SAPS II, the MPM_0 II, and the MPM_{24} II models on a large data set from the Dutch NICE registry. To put the historical external validation studies of the four prognostic models into perspective, we also investigated the influence of sample size on the validation results. To this end, the validation process was repeated with smaller data sets that were randomly drawn from the NICE registry.

2. Methods

2.1. Data

In 1996, the Dutch NICE foundation has started the (voluntary) registration of data of admissions to Dutch ICUs. The NICE registry database contains for each ICU admission 108 demographic, diagnostic, and physiologic variables collected within the first 24 hours of ICU admission and outcome data, such as length of stay on ICU and in-hospital mortality.[1] Data collected include all raw data values necessary to calculate the original SAPS II [3], APACHE II [4], MPM_0 II, and MPM_{24} II [5] mortality probabilities. APACHE II, SAPS II, MPM_0 II, and MPM_{24} II mortality probabilities are calculated in the national database at the NICE data coordinating center. Stringent measures are taken to control the data quality and uniformity of data collection procedures in the participating ICUs [18,19].

[1] http://www.stichting-nice.org

The data set used in this study consisted of data from 83,824 admissions to 29 Dutch ICUs between January 1, 1999 and December 31, 2003 registered in the NICE database. The developers of the APACHE II, SAPS II, MPM_0 II, and MPM_{24} II models have defined criteria for populations on which the models can be applied. We combined the criteria of all four models to obtain one data set that satisfied all criteria. According to the combined criteria we excluded patients aged <18, patients with an ICU length of stay <8 hours, acute coronary care and cardiac surgery patients, burn patients, readmitted patients, patients with missing severity-of-illness scores, and patients with missing (hospital) survival status. The characteristics of the remaining data set were compared to those used to develop the original APACHE II, SAPS II, MPM_0 II, and MPM_{24} II models.

2.2. Validation measures

2.2.1. Discrimination

The term *discrimination* refers to a model's ability to distinguish survivors from nonsurvivors. As a measure of discrimination we calculated the area under the ROC Curve [10]. This Area under the Curve (AUC; sometimes called C-index) is a normalized Mann–Whitney U statistic applied to the predictions by the model, grouped by observed outcomes. It 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. Under the assumption that the distribution of AUCs is approximately Normal, we can compute the standard error of an estimated AUC [10].

For each pair of models (six in total), the difference in AUC was statistically tested with the nonparametric method of DeLong et al. [20]. The main problem in testing the difference between two AUC values that were computed on the same data set is that these values are highly correlated. The method of DeLong et al. solves this problem by estimating the correlation between the two values using the theory of generalized U statistics.

The AUC of a model depends only on the order of observations induced by its predictions and provides no indication of how close, on average, the predicted probabilities are to the observed outcomes. To take this aspect of a model's performance into account, we have to look at the *accuracy* and *calibration* of a model.

2.2.2. Accuracy

Accuracy refers to the difference between predicted risks and observed outcomes at the level of individuals. In this study, we applied the Brier inaccuracy score. The (*mean*) *Brier inaccuracy score*, also known as *mean squared error* or *mean probability score* [21,22], is calculated as