Comments on Practical experiences on the necessity of external validation by I. R. König, J. D. Malley, C. Weimar, H.-C. Diener and A. Ziegler, Statistics in Medicine, DOI: 10.1002/sim.3069

Björk, Jonas; Green, Michael; Ekelund, Ulf

Published in:
Statistics in Medicine

DOI:
10.1002/sim.3168

2008

LETTER TO THE EDITOR

Comments on 'Practical experiences on the necessity of external validation'

by I. R. König, J. D. Malley, C. Weimar, H.-C. Diener and A. Ziegler, *Statistics in Medicine*, DOI: 10.1002/sim.3069

From: *Jonas Björk*[*,†,1], *Michael Green*[2] *and Ulf Ekelund*[3]
[1]*Competence Center for Clinical Research, Lund University Hospital, Lund, Sweden*
[2]*Department of Theoretical Physics, Lund University, Lund, Sweden*
[3]*Department of Clinical Sciences, Section for Emergency Medicine, Lund University Hospital, Lund, Sweden*

In a very interesting empirical study, König *et al.* investigated the ability of two cross validation (CV) methods, tenfold CV and leave-one-center-out CV, to predict the generalizability of prognostic models [1]. Three different prognostic models, based on logistic regression, support vector machines and random forests, were established and validated. The goal of the prognostic models was to predict accurately the functional independence status of patients 100 days after an acute ischemic stroke. Data from the German Stroke Data bank were used.

The CV methods, which operate on training data, generally overestimated the temporal and geographic transportability of the prognostic models. No important differences in the distribution of the prognostic factors between the training set and the temporal and external validation sets were discerned that could explain the overoptimistic results of the CV methods. However, distributional data for the outcome measure were not presented or discussed. The outcome, the functional status of each patient after 100 days, was measured by Barthel index (BI), which is a continuous measure between 0 (total functional dependence) and 100 (total functional independence). In the prediction models, BI$\geqslant$95 was used as a binary outcome measure. This, however, imposes a sharp cut-off for something that is indeed continuous. The discrimination problem can be expected to be much easier if the underlying continuous outcome for most patients are either far below or far above the chosen cut-off, compared with if many patients have values near the cut-off [2]. It is therefore recommended that detailed distributional data of continuous measures of the severity of disease (*here* BI) are presented [3], both for the training and validation sets.

When comparing the two CV methods, the authors draw the intuitively appealing conclusion that the leave-one-center-out approach is probably better than the traditional tenfold CV in predicting geographic transportability to centers not included in the training set. However, there does not seem to be a firm support for this conclusion in their data, given that there is a fairly substantial overlap in the confidence intervals for the estimated accuracy of the two CV methods (see the results for logistic regression in Figure 2 in the original article). It should be stressed that the leave-one-center-out approach can be expected to yield uncertain estimates of the geographic transportability when the number of centers in the training set is limited.

*Correspondence to: Jonas Björk, Competence Center for Clinical Research, Lund University Hospital, Barngatan 2, SE-221 85 Lund, Sweden.
†E-mail: jonas.bjork@skane.se

The work by König *et al.* stresses the importance of external validation. In many applications, prognostic models have to be optimized for each new target population before implementation [4]. After such optimization, there is a need to estimate the temporal transportability of the redefined model within the new target population. The results of König *et al.* are a bit reassuring in that respect, since they show that CV methods may predict temporal transportability fairly well, at least for their particular application and the investigated time period. It is not obvious, however, that this is generally true for other disease groups as well, where changes in, e.g., treatment practices or diagnostic criteria may have occurred over time.

REFERENCES

1. König IR, Malley JD, Weimar C, Diener HC, Ziegler A. Practical experiences on the necessity of external validation. *Statistics in Medicine* 2007; DOI: 10.1002/sim.3069.
2. Levey AS, Coresh J, Greene T, Stevens LA, Zhang YL, Hendriksen S, Kusek JW, Van Lente F. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Annals of Internal Medicine* 2006; **145**:247–254.
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clinical Chemistry* 2003; **49**:1–6.
4. Kennedy RL, Burton AM, Fraser HS, McStay LN, Harrison RF. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. *European Heart Journal* 1996; **17**:1181–1191.

Published online 2 January 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.3168

AUTHOR'S REPLY

In their comments on our recent publication [1], Björk *et al*. [2] raise two important discussion points. We welcome their comments and concerns as well as the opportunity to reply adequately. The questions they raise are important and are part of a large and continuing discussion of how best to choose outcome measures, and then how to most efficiently validate any predictive method. The first issue concerns the distribution of the outcome measure we employed. To be specific, we relied on the Barthel index (BI) after 100 days and dichotomized this to classify patients as being completely restituted after the stroke (BI$\geqslant$95) *versus* incompletely restituted or deceased (BI$<$95). The authors are concerned about this dichotomization of an allegedly continuous measure. As mentioned above, such concerns are long-standing and are not limited to the study at hand, see for example [3]. Intuitively, it is clear that the dichotomization of scales could in principle reduce outcome information and may limit a scale's ability to detect a significant shift in disability [4, 5].

There were three reasons for our proceeding as we did. Firstly, the BI is not continuous in nature. In fact, it consists of 20 items for which 0 and 5, and in some cases 10 and 15 points can be scored [6]. Thus, it is an ordinal scale with 21 levels, and the gaps between the scores are not necessarily equal [7]. Based on this, categorical analyses have been assumed to be appropriate [7]. Our second reason for choosing a dichotomization was that for the conductance of clinical trials on stroke this is clearly recommended. It is difficult to understand, generally, how a continuous outcome by itself should be uniformly recommended as a replacement or sound alternative to a clinically mandated, yes/no classification rule for patient prognosis. Indeed, in the Points to

