Flaws in the development of the CHESS score

G. S. Collinsa and Y. Le Manachb

aCentre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK, and bDepartments of Anesthesia and Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, Population Health Research Institute, McMaster University and the Perioperative Research Group, Hamilton, ON, Canada

Correspondence: G. S. Collins, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Windmill Road, Oxford OX3 7LD, UK (tel./fax: +44 (0) 1865 223460; e-mail: gary.collins@csm.ox.ac.uk).

Keywords: calibration, overfitting, prediction model, reporting, TRIPOD

doi:10.1111/ene.13036

Received: 18 March 2016
Accepted: 22 March 2016

Re: The CHESS score: a simple tool for early prediction of shunt dependency after aneurysmal subarachnoid hemorrhage

We read with interest the recent paper by Jabbarli and colleagues describing the development of the CHESS score to predict shunt dependency in patients following aneurysmal subarachnoid haemorrhage [1]. However, numerous deficiencies in the statistical analysis whilst developing the model, lack of internal validation and an inappropriately small validation dataset cast serious doubts on the usefulness of the model. Here we constructively summarize some of the more major concerns.

First, the sample size is small in relation to the number of variables analysed: 75 outcomes (shunt replacements) and 23 variables, leading to an events-per-variable of 3, much lower than the recommended value of 10, and therefore the risk of overfitting is a concern [2]. Selecting variables based on their unadjusted association with the outcome is a flawed approach [2], let alone basing this on calculating the Spearman correlation. Furthermore, analysing continuous variables as both continuous and dichotomized (and for age also categorized) is highly questionable and should be avoided [3]. Internal validation (which is absent from this study) using recommended methods such as bootstrapping to quantify and adjust for overfitting should be considered given all the aforementioned analysis deficiencies.

Whilst validation in separate data is a suitable test of a model, the sample size of the data used to validate the CHESS score is far too small (30 patients; number of outcomes not reported) to allow a meaningful evaluation (sample size considerations for validation require a minimum of 100 outcome events [4]). The Hosmer–Lemeshow test is also inappropriate to assess calibration as it quantifies neither direction nor magnitude of any miscalibration. The recommended approach would be to present calibration plots [2].

We recommend the authors and indeed other investigators developing a risk score to consult the TRIPOD statement (www.tripod-statement.org) for key information to report when describing its development and validation, so that readers have the minimal information required to judge the quality of the study and therefore whether to use the model [5]. The accompanying TRIPOD explanation and elaboration paper also discusses various methodological considerations that investigators should consider when developing and validating a risk score.

Disclosure of conflicts of interest

The authors declare no financial or other conflicts of interest.

References