Design flaws in EuroSCORE II

Gary S. Collins* and Douglas G. Altman

Centre for Statistics in Medicine, Wolfson College Annexe, University of Oxford, Oxford, UK

* Corresponding author. Centre for Statistics in Medicine, Wolfson College Annexe, University of Oxford, Linton Road, Oxford OX2 6UD, UK. Tel: +44-1865-284418; fax: +44-1865-284424; e-mail: gary.collins@csm.ox.ac.uk (G.S. Collins).

Received 28 August 2012; accepted 24 September 2012

Keywords: EuroSCORE • Design • Risk prediction • Validation

We recently read with great interest and anticipation the development of the EuroSCORE II. However, we have methodological concerns relating to how EuroSCORE II was developed and evaluated [1] and subsequent comments made in an editorial by the authors responding to recent criticisms [2]. We focus on the strategy and study design used to derive and evaluate the model. The authors randomly split the EuroSCORE data into two data sets, one to derive EuroSCORE II and the other to evaluate its performance. This approach, while unfortunately widespread, has been shown to be ineffective and should be avoided in favour of more methodologically sound methods, such as bootstrapping [3]. One reason is that the reliability of a model will be diminished if it is derived from a reduced data set. For small data sets overfitting is a concern, producing unstable and inaccurate risk scores.

Delving deeper, evaluating the performance of a risk score is ultimately all that matters, and how it was derived is generally only of interest to gain insight when a risk score fails to work. Strategies to evaluate performance are broadly referred to as internal, temporal or external validation [4]. Internal validation involves using the original data (including data splitting) for both the development and the initial validation of the risk score [5] (the strategy used to develop the EuroSCORE II [1]). Temporal validation uses data collected in the same places at a different time. External validation uses a separate, independent data set collected by different investigators at different centres, possibly in a different time period. In a recent correspondence responding to criticisms of the validation of EuroSCORE II, the authors authoritatively, yet falsely, asserted that randomly splitting the data into the development and validation data sets constituted the strongest test of EuroSCORE II [2]. That claim is incorrect and ignores a large body of methodological literature on the design and evaluation of risk scores [4, 6]. Among many methodological concerns, random splitting large data sets just create two closely similar (apart from chance) data sets and is hardly a tough test [4]. Importantly, it does not constitute an external validation of the EuroSCORE II.

Furthermore, during the evaluation of EuroSCORE II, the authors conducted a 10-fold cross-validation to, in the authors own words, ‘assess the validity of the model’ [1]. Cross-validation is an approach (similar to bootstrapping), whereby all the data are to develop and validate the risk score [6]. Cross-validation is a stronger design than the random split-sample approach, but these are distinct, mutually exclusive approaches (along with bootstrapping) to develop and evaluate a risk score and not to be used side-by-side. Thus, the results from the cross-validation reported in the development of EuroSCORE II [1] should not be taken as providing robustness to their results.

Validation studies are important because the performance of prediction models tends to be poorer when applied to new individuals than in the sample from which it was developed. Truly external validation studies, using independently collected data are needed to evaluate the potential clinical usefulness of the EuroSCORE II.

REFERENCES