Towards improved risk scores: the quest for the grail continues

Bernard Iung1* and Alec Vahanian2

1 Cardiology Department, Bichat Hospital, AP-HP, 46 rue Henri Huchard, 75018 Paris, France; and 2 University Paris Diderot, Paris, France

This editorial refers to ‘Does EuroSCORE II perform better than its original versions? A multicentre validation study’, by F. Barili et al., on page 22

Barili et al. have reported the first external validation of the recently published EuroSCORE II scoring system.1

A new scoring system was needed to replace the original logistic EuroSCORE, not only because it was elaborated >15 years ago, but, more importantly, because a number of papers showed that the EuroSCORE system was poorly calibrated when applied to contemporary data sets. This means that there were discrepancies between predicted and observed operative mortality, with a trend to overestimate the operative risk. The EuroSCORE II system was recently shown to achieve a similar discrimination to the original EuroSCORE, but to be better calibrated.2 In the initial study, validation was performed using a different data set from the one used to elaborate the scoring system, but this was not an independent validation since the validation and derivation samples were produced from the same data collection and from the same centres.2 The main conclusions of the external validation are that the EuroSCORE II system achieves a similar discrimination to EuroSCORE, but not a better calibration.1

The strength of the study by Barili et al. is that the authors have performed an external validation in a large data set prospectively collected from different types of hospitals. The discriminatory properties of the EuroSCORE system have never been a major source of concern. The area under the receiver operating characteristic (ROC) curve was 0.82 for EuroSCORE II in the external validation and 0.81 in the internal validation. These values are close to the discrimination obtained with the original EuroSCORE and with the Society for Thoracic Surgeons (STS) score.3,4 Therefore, the fact that discrimination does not appear to be improved with the EuroSCORE II system is not a drawback in itself.

On the other hand, external validation leads Barili et al. to conclude that the calibration properties of EuroSCORE II do not seem to be significantly improved as compared with EuroSCORE. This statement seems to go too far given the results of the external validation. The statistical significance of goodness-of-fit tests, such as the Hosmer–Lemeshow test, actually means that there are significant differences between the numbers of observed and predicted deaths. However, discrepancies between observed and predicted deaths in a particular subgroup of patients, such as high-risk patients, may lead to statistically significant differences of the overall test, in particular when using large data sets with a large number of events. Therefore, the P-value of 0.001 in the external validation does not summarize, per se, all of the calibration properties of the EuroSCORE II system. Rather than overall statistical tests, the graphical representation of observed vs. predicted mortality is of great interest to assess the usefulness and limitations of a scoring system in clinical practice. As pointed out by the authors, the graphical representation shows that EuroSCORE constantly overestimates the operative mortality, whatever the risk considered. This is not fully in accordance with previous analyses which showed that the miscalibration of EuroSCORE was mainly observed in high-risk patients.3–7 On the other hand, there is a good agreement between observed and predicted operative mortality with the EuroSCORE II system when considering patients with a predicted operative mortality of <30%. This is of considerable interest since patients with a predicted operative mortality of <30% account for most patients who are operated on. Unfortunately, the distribution of operative risk and the detailed description of the population are not provided in this validation sample. In the EuroSCORE II population, observed operative mortality was 18% (280/1595) in the highest risk decile.2 Therefore, the external validation suggests that EuroSCORE II leads to a much more accurate estimation of operative mortality than EuroSCORE for the vast majority of patients.

Nevertheless, a poor calibration and an overestimation of the operative risk are still observed with EuroSCORE II in high-risk patients.1 This has an important impact in practice since risk assessment has recently gained importance in this particular subgroup with the development of percutaneous treatment of valvular heart disease.

The opinions expressed in this article are not necessarily those of the Editors of the European Heart Journal or of the European Society of Cardiology.

* Corresponding author. Tel: +33 1 40 25 67 60, Fax: +33 1 40 25 67 32, Email: bernard.iung@bch.aphp.fr
1 doi:10.1093/eurheartj/ehs342
© The European Society of Cardiology 2012. All rights reserved. For permissions please email: journals.permissions@oup.com
When transcatheter aortic valve implantation (TAVI) became widely available, a European consensus document mentioned the values of 20% for EuroSCORE and 10% for the STS score as thresholds above which TAVI should be considered. This paper also stressed that risk scores should be only one of the components of the decision, which relies more on clinical judgement than on score values. There is now considerable experience with TAVI from large registries and randomized trials. During the same period, a number of papers have analysed contemporary operative mortality and risk factors of surgical aortic valve replacement. This reinforces the former statement concerning the limitations of risk scores. For example, in the two cohorts of the Partner trial, the values of EuroSCORE and STS scores were nearly similar, although patients were considered inoperable in cohort B, and at high risk for surgery, but operable, in cohort A. The discrepancies between observed and predicted operative mortality, in particular when using the logistic EuroSCORE, have also been consistently reported in analyses of surgical databases focusing on high-risk patients.

The value and limitations of risk scores when applied to patients with valvular heart disease have recently been reviewed in a position paper from the ESC Working Group on valvular disease. The limitations in the performance of risk scores in high-risk patients with valvular disease may be due to a number of factors. Population characteristics are one of these factors, since high-risk patients often account for only a small proportion of databases from which risk scores are elaborated and validated. Risks inherent to techniques are subject to changes over time. The choice of variables and variable coding should find a difficult compromise between the completeness of patient characteristics analysed and the ease of use of the scoring system. Conditions that are relative or absolute contraindications to surgery are, of course, very rare in surgical series, and their prognostic value is therefore difficult to determine. Finally, indices of cognitive or functional performance are not included in current risk scores, although they have an impact on decision making for intervention in the elderly. The EuroSCORE II system adequately addresses some of these factors. The database from which EuroSCORE II was derived reflects contemporary populations and practices. It is more diverse than for EuroSCORE and includes 46% of cases of valvular surgery. Changes in variable coding allow for a better estimation of the impact of renal function and the type of surgical procedure. An important challenge in the elaboration of a risk score is the choice of variables. The inclusion of variables related to frailty seems to be of interest in high-risk patients. This should rely on validated and relatively simple indices such as the Katz index or the Instrumental Activities of Daily Living score. However, the heterogeneity of the population of high-risk patients makes it unlikely that the contribution of each co-morbidity or functional impairment could be accurately ascertained. In the majority of patients, who are at low to intermediate risk, risk assessment should rely on a user-friendly score based on a limited number of variables which are easy to collect, such as in EuroSCORE II. The possibility of obtaining a similar performance with a smaller number of variables is suggested in the paper by Barili et al.

These observed and intrinsic limitations of risk factors in high-risk patients are the reason leading to the recent ESC/EACTS guidelines on the management of valvular disease recommending that risk assessment should not come from a single number of a risk score, but from an overall assessment through a multidisciplinary approach, involving the collaboration of the Heart Team. This is particularly emphasized for the choice between surgery and TAVI in high-risk patients. Of course, a scoring system should be applied only for the type of procedure from which it was elaborated. Unsurprisingly, the performance of the EuroSCORE system is poor when applied to TAVI, and specific risks scores are needed in this field.

Multivariate scoring systems are presently the only means of reducing the subjectivity of risk estimation, and their use should thus be encouraged. It is necessary to continue to perform external validation in different populations to refine the assessment of their performance in different subgroups of patients. Discrimination and calibration properties of EuroSCORE II make it a useful tool in decision making for many candidates for surgery. It is, however, important to keep in mind the limitations of current risk scores in high-risk patients, in whom predicted values should be integrated into, but should not be a substitute for, clinical judgement.

Conflict of interest: B.I. has received consultant fees from Servier, Boehringer Ingelheim, Bayer, Valttech, and Abbott, and speaker’s fees from Edwards Lifesciences, St. Jude Medical, and Sanofi-Aventis. A.V. is a member of Advisory Board for Medtronic, Abbott, Valttech, and Boehringer Ingelheim, and has received speaker’s fees from Edwards Lifesciences and Siemens.

References


A 46-year-old woman was admitted to hospital due to dyspnoea. Chest X-ray, serological parameters, and spirometry ruled out pulmonary reasons or infection as its cause.

Transthoracic echocardiography revealed a moderately impaired left ventricular (LV) contractility (ejection fraction 42% calculated according to Simpson’s rule); there were no relevant valvular pathologies. In the apex of the LV, prominent trabeculae were detected (Panel A); colour Doppler showed perfusion of the intertrabecular spaces from the LV cavity. This raised the suspicion of non-compaction cardiomyopathy (NCCM). Coronary angiography was performed, which ruled out coronary artery disease. Surprisingly, after injection of contrast agent (CA) into the left coronary artery (LCA), CA was rapidly detected within the LV, indicating diffuse fistulae from the LCA to the LV (Panel C: angiogram of the LCA, Panel D: beginning inflow of CA from the LCA into the LV, Panel E: opacification of the LV by the CA, Panel F: magnification of the marked area showing CA passed over from the LCA to the LV). For further diagnosis, magnetic resonance tomography was performed. This excluded a myocarditis, but confirmed non-compaction in the region, where the coronary fistulae were detected (Panel B).

Non-compaction cardiomyopathy is a rare cause of congestive heart failure, with an incidence of 0.05–0.25% in the whole population. During embryonic development, disturbed compression of the trabeculated myocardium leads to non-compaction with impaired contractility. Theoretically, this process may lead to coronary fistulae, too. Nevertheless, the coincidence of non-compaction cardiomyopathy and coronary fistulae has been described only a few times before.