We are pleased that EuroSCORE II [1] continues to generate interest and are happy to respond to concerns about its development and evaluation [2]. We find very little to disagree with in the points raised by Collins and Altman [3], but are not convinced that their suggested alternative analysis would have resulted in materially different results. Our approach to the analysis was traditional (split data design) and lacked the efficiency of sampling-based methods, such as bootstrapping. Concerns regarding split data designs usually affect much smaller datasets, in which overfitting is a more likely. Our development dataset contained over 16,828 cases and, depending on definitions, between 20 and 30 events per variable, so that standard errors around the estimates were sufficiently small. Careful development of risk models from large datasets using bootstrapping is more computer intensive and would have achieved only a marginal increase in precision. The logistic decisions we made included trade-offs of time, resources, validity and precision of estimates.

Our main reason for using the split data design was to have an independent dataset on which to test the EuroSCORE II. This is an internal validation method and, since the data were split randomly, validation cases will be similar to those from which the score was derived, giving optimistic results. Of course, external validation in different related populations is a stronger test. Scores should always be externally validated, and we have consistently urged other groups to conduct and publish their own analyses. The most valuable are those which are multi-institutional and address the global cardiac surgery population for which the model is intended. In this respect, we agree absolutely with Collins and Altman. However, we must clarify that the statement they quote from our discussion completed a paragraph on the validation we had performed. We had not intended to suggest that any internal validation method was the strongest possible, just that this was the strongest test we had conducted. We apologize if this led to any confusion.

We also clarify that 10-fold cross-validation was used as an additional internal validity analysis. Briefly, the model was refit on 10 × 90% samples using the independent variables identified from the development set, and measures of model fit (Hosmer–Lemeshow) and discrimination (concordance statistic c) were estimated for each model. It was not used to develop a different model.

External validation of the EuroSCORE II is emerging. Grant et al. [4] have shown that it performs well overall in 23,740 UK cardiac surgery cases, although it significantly overestimated risk for isolated coronary artery bypass grafting patients (observed mortality 1.5% and predicted mortality 2.1%). These published results allowed us to calculate Altman and Royston’s PSEP (index of separation) statistic [5] showing the separation of the score for the groups with the best and worst predicted outcomes. We found that the development and validation sets have similar PSEP (16 and 15.9%) and, as expected, they are higher than for the external dataset (13.7%), but not by much. We are therefore confident that alternative methods, though valid, would not have produced a discernibly different EuroSCORE II.

REFERENCES