Applying and evaluating risk models

Samer A.M. Nashef*

Papworth Hospital, Cambridge, UK

* Corresponding author. Papworth Hospital, Cambridge CB23 3RE, UK. Tel: +44-1480-364299; fax: +44-1480-364744; e-mail: sam.nashef@papworth.nhs.uk (S.A.M. Nashef).

Keywords: Risk modelling • EuroSCORE • Mortality

Diem et al. [1] compared two risk-adjusted mortality models in cardiac surgery: EuroSCORE (both the additive and logistic versions) and the Cardiac Anesthesia Risk Score (CARE) in a population of 3818 cardiac surgery patients from the University of Ottawa Heart Institute in Canada. They found not only that the logistic EuroSCORE has the best discrimination (area under the receiver operating characteristic (ROC) curve of 0.84), but also that it has the worst calibration with predicted mortality of 7.72% (actual 3.25%). CARE has good discrimination (area under the ROC curve of 0.79) and good calibration (predicted mortality of 3.38%). They conclude that CARE is a good model with robust calibration 10 years after its development and discrimination that is acceptable although not as good as the logistic EuroSCORE. The CARE model also has the advantage of relatively few risk factors (6 versus 17). This is of course a valid conclusion, but it must be tempered by a few other considerations. No risk model is perfect and for that reason any model is necessarily a compromise between what is feasible and what is sensible. Risk modelling is carried out for reasons which fall broadly into two categories.

The first is to assess risk for the patient. In this area, the more comprehensive the risk model is, the more accurate is the risk assessment. It can of course be argued that all models fail in the individual patient: a model which predicts a mortality of 3% for an individual patient will always be wrong, as the patient will be either dead or alive after surgery, and never ‘3% dead’. Despite that, knowledge that a particular procedure in hundreds of similar patients will have 3% mortality remains a useful element in the information guiding the clinical decision. All such decisions depend on weighing the benefit against the likely risk, and this is true when a surgeon offers a treatment to a patient and when the patient gives informed consent to receiving it. Indeed, when operation is being contemplated on prognostic grounds alone in an asymptomatic patient, then evaluating the risk of surgery becomes imperative: we must not offer an operation to an asymptomatic patient if surgery carries a greater risk than conservative treatment.

The second is to evaluate care. A risk model offers a standard expected outcome against which the actual outcome can be measured. If the risk model says that your predicted mortality should be 5% for a certain group of patients with a particular risk profile, and your actual mortality is 1%, then you are doing well. Dividing actual mortality by your predicted mortality is termed the ‘risk-adjusted mortality ratio’ or RAMR. This can be given with a 95% confidence interval and is probably the most useful single measure of the performance of a cardiac surgical unit. RAMR can be used to assess quality, record improvement and underperformance and to compare units and surgeons on a ‘level-playing field’. When a model is used to evaluate care for a large group of patients, the importance of rare and obscure risk factors is reduced, and it can be shown that simple risk models such as CARE (or even simpler ones based on two or three factors) can achieve reasonable quality assessment. Cardiac surgeons, however, are busy people and may prefer a single model for all purposes, hence compromise models such as EuroSCORE.

Models such as CARE are nevertheless useful and easy to apply for evaluating quality. However, such evaluation is a highly sensitive area, on which reputations and careers may depend, and it is important in this field that definitions of risk are neither subjective nor open to observer interpretation. In this context, the CARE model may be vulnerable, as the definitions of what exactly constitutes ‘uncontrolled medical problem’, ‘complex surgery’ and so forth are not crystal clear, and the model requires the use of clinical judgement in the determination of risk.

As this paper eloquently shows, the validation of a risk model depends on the assessment of two features: calibration and discrimination. Calibration is the accuracy of the model for predicting risk in a group of patients; in other words, if the model says that mortality in a thousand patients is likely to be 5%, and actual mortality is 5% or close to it, then the model is well calibrated. Discrimination is the ability of the model to distinguish between low-risk and high-risk patients. In other words, if most of the deaths occur in patients that the model correctly identifies as high risk, then the model has good discrimination, but if most deaths occur in patients that the model actually identifies as low risk, then there is poor discrimination. We measure discrimination using a statistic called the ‘area under the ROC curve’, sometimes also called the c-statistic or the c-index. If the area under the ROC curve is 0.5, then the model does not discriminate at all. Good discrimination begins at 0.7 and rarely exceeds 0.85. If the area under the ROC is 1.0, the model becomes a crystal ball which tells us the future (an impossible task).

A risk model may have good calibration but poor discrimination and vice versa. Discrimination is more important than calibration. A model can be recalibrated or adjusted as practice
improves, but if the model is built on the wrong risk factors, its discrimination cannot be improved.

Despite all of the above arguments, there is now a strong case, made by this paper and others, for reviewing the EuroSCORE model in order primarily to improve calibration, but also to attempt even better discrimination in the light of studies which have proved a number of areas for potential improvement. Data from more than 20,000 patients have been collected from more than 150 centres in 43 countries in order to renew the model along the lines described above.

The new model was presented at the EACTS meeting in Lisbon in 2011.

REFERENCE