Complexity and Severity Scores in Cardiac Surgery. Uses and Limitations

José M. Cortina Romero

Servicio de Cirugía Cardíaca, Hospital 12 de Octubre, Madrid, Spain.

In recent years, the use of predictive models for estimating the risk of mortality associated with heart surgery, and in particular coronary revascularization surgery, has become common practice for heart surgeons and cardiologists. This is true to such an extent that the use of these models receives a class IIa recommendation (evidence grade C) in the 2004 clinical practice guides of the AHA/ACC.[1] This recommendation refers explicitly to the use of predictive models for the preoperative estimation of the above-mentioned risk, a practice that helps doctors and patients weigh up the risks and benefits of the procedure proposed. However, these systems have other uses. While they certainly provide preoperative risk estimates for individual patients—perhaps the most intuitive and therefore the most common use made of them by clinicians—it should be remembered that they were originally developed for making overall estimates with respect to whole series of patients. To explain this difference in use, the origin of the different models must be examined. Systems for predicting and adjusting the risk associated with heart surgery have existed since the time of the CASS.[2] However, the real take-off in their use, as we understand it today, came after the raw mortality results for hospitals that operated on MEDICARE patients were published by the Health Care Financing Administration (HCFA) in March 1986. This led to the Society of Thoracic Surgeons (STS) of the USA taking the position[3] that the use of mortality data without appropriate adjustment for risk factors was inappropriate and incorrect. From that moment, systems began to appear that weighted results in terms of severity of disease and the existence of associated morbidity.

SEE ARTICLE ON PAGES 515-22

Correspondence: Dr. J.M. Cortina Romero.
Servicio de Cirugía Cardíaca. Hospital 12 de Octubre.
Avda. Córdoba, s/n. 28041 Madrid. España.
E-mail: jcortina.hdoc@salud.madrid.org

METHODOLOGY

It is not the aim of this commentary to describe the development, assessment and validation of the predictive models from which different scores[4] are derived. However, it should be remembered that the maximum methodological robustness is required in their construction.[5] Briefly, the first step in the development of such models requires the precise definition of the variable under examination, normally in-hospital death, followed by an analysis of the factors that might influence this. Surprisingly, the precise definition of such variables is one of the most difficult tasks. Even death can be defined in several ways. Indeed, it was the imprecise definition of certain variables that weakened one of the pioneering models.[6] Another point of controversy is the number of variables that a model should take into account. From the standpoint of everyday clinical practice, it might appear that the more variables included, the more likely the model will reflect reality. While this is essentially true, it only applies when models are used for estimating risks for individual patients. It has been shown that statistically robust models useful for making predictions regarding whole series of patients can be developed using only a small number of essential or "central" clinical variables. Adding new variables beyond a certain number only marginally increases their predictive power. It should be remembered that the prospective use of these models requires every single patient be scored —without exception. Clearly this is easier when there are fewer variables and when these are precisely defined. It is also important to take into account the greater predictive power of models based on clinical rather than administrative data.

Once a model has been constructed it needs to be validated; this involves a series of steps to determine

whether the model is reliable and robust. The normal interpretation of what validation entails (in terms of the everyday use of a model) refers to the validation of its predictive power; but this is only one of several important points that need to be taken into account. This validation of predictive power involves two well known factors: calibration and discriminating power. Calibration evaluates a model with respect to its capacity to predict overall mortality, as well as mortality with respect to different risk strata. Discriminating power,[7] however, is a measure of how well a model predicts a certain result; this generally depends on the area under the ROC curve. Excellent discrimination is reflected by values of greater than 0.97. The range 0.93-0.96 represents very good discrimination, 0.75-0.92 represents good discrimination, and anything below 0.75 represents deficient discriminating power.

USES AND LIMITATIONS

These models can be used for 3 different, although related, purposes: for estimating the risk for a single patient, as descriptors of the case-mix of patient populations, and as quality control and management tools. The best model to use will depend upon the task at hand.

Use With Individual Patients

It should be remembered that these scores were not developed for use with single patients. Although they have good discriminating power, it can never be as high as 1. Therefore, the use of these models with any particular patient can only be orientative. A determined risk can be estimated, but the final result never predicted. In other words, models with good predictive power may show that there will be 5 deaths among 100 patients—but they can never predict which 5 patients will die.

As recommended in the AHA/ACC guidelines, these models can be helpful when deciding upon the best therapeutic course to follow. The divergence between subjective estimates of risk and those provided by these scores (with respect to an individual patient) is surprising. For individual estimations of risk, the most logical recommendation is that the model used be based on the experience of the center where therapy is to be provided. However, while there are groups that have developed their own predictive model, such proliferation of modeling does not occur.

For individual patients, logistic models that contemplate the greatest number of variables should be used. These models should take into account the entire clinical profile of the patient. The current Bernstein and Parsonnet model[8] or that posted on the STS website (www.sts.org) approximate to these requirements.

Giving medical advice with respect to a high risk procedure is difficult. True it is that patients at the highest risk are those who most benefit from such procedures if they survive, but it is also true that there are levels of risk that, in practice, mean the chances of survival are minimal or even nil. Giving advice on what therapy to follow can be very complex in such cases.

Use as Descriptors of the Case-Mix of Populations

One of the virtues of these types of score systems is that they summarize in a single number the clinical profile of individual patients, including data on the severity of the main disease and its associated pathologies. This allows a simple evaluation of the overall characteristics of a population of patients—the case-mix—to be made. In turn, this allows different populations (groups, hospital populations, even different countries) to be compared. For the same reason, changes over time in the same institution can be followed. The change in the case-mix reported by García Fuster et al[9] (who used these methods) showed a significant increase in the disease severity of their populations between the first and third 3-year periods, followed by stabilization (although with a small, non-significant deterioration).

For the use of these scores as descriptors of populations, it is clear that data do not have to come from the current population. However, the definition of these scores cannot be changed at will since this would render comparisons impossible.

The use of these scores in this area has two particularly important limitations. Firstly, it has been shown in the State of New York (following the publication of mortality results by center and surgeon) that there is a possible tendency to artificially overload the case-mix, especially if imprecisely-defined variables are used. Logically, the resulting score would not be that which truly corresponded to a strict use of the model. The only way to overcome this is the exclusive use of precise, unquestionable variables, the use of the scores by external agents, and the systematic auditing of the information-gathering process. Secondly, these scores cannot detect variations in the criteria for the indication of surgery either between groups or within the same group. Thus, variations in the case-mix could translate into differences in the selection of patients without there necessarily being any differences in the characteristics of the populations requiring attention.

Quality Control

It might be claimed that this is the most basic use that can be made of risk scores. The main aim is that they estimate the results that might be expected depending on the type of population treated. If, as usual,