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,

the score is understood as an individual estimate of the risk of death for each patient, then the risk of death for a series of patients is the mean of the individual risks. The comparison of the observed mortality and the estimated mean risk provides a result that allows different series of patients to be compared under comparable conditions. The most intuitive way of handling these data is to calculate the ratio between the observed mortality and the mean estimated risk of mortality. Values <1 indicate results better than those expected, whereas those >1 would be worse than expected. All these steps must be accompanied by appropriate statistical methods for measuring dispersion.

The use of these models in quality control gives rise to a number of difficulties. The most common question is which score to use. The first condition for the use of a model should be that it is robust. It is also recommended that it be constructed using databases that refer to patients similar to those to be treated. Perhaps the most recommendable score at the present time is the EuroSCORE.[10] This score was constructed using a European database including 20 000 patients who underwent treatment in the last trimester of 1995. Spain contributed more than 2000 cases (involving more than 20 hospitals). The EuroSCORE has been sufficiently validated in Europe and shows good calibration and discrimination. The same has been reported when it has been used in populations with important demographic differences, for example those of North America. Although it is reasonable to think that this model would not work well for OPCAB patients, the opposite appears to be true[11]; in fact it shows good calibration and discriminating power.

The ideal predictive model is one based on an extensive database that is updated over time and that reflects the day to day changes introduced into clinical practice. The STS model, which is now about 10 years old, meets these criteria. In Europe, the foundations are being laid for a similar model, involving the creation of a European database by the European Association for Cardiothoracic Surgery (now in its initial phases).

The use of these types of score in quality control has a number of serious limitations, and errors of interpretation are common. Firstly, although these tools are statistically sophisticated, they are not infallible. Certainly they take into account only a small number of the many known patient and health service variables that can influence the final result (there may also be unknown variables). Thus, the conclusions that can be drawn after their use have to be carefully qualified. Another important limitation has already been outlined: these scores habitually reflect a snapshot of current clinical practice. The work of García Fuster et al[9] is a clear example of this. The use of the EuroSCORE indicates a favorable progression in terms of results achieved. According to these authors, this reflect an improvement in the service provided. However, it is quite possible that had a model been used that reflected current practice, the improvement seen in the results would not have been the same.

An error is often made due to "circularity." It is quite common to use scores such as the EuroSCORE, and then to compare their discriminating power and calibration with a model based on a center's own experience. Normally, the discriminating power of an in-house model is much greater than that of external models. Statistically this might be expected, but from the point of view of quality control it is invalid. It is easy to understand that the use of data from a group with very high mortality rates might lead to the development of a predictive model with very good calibration and discriminating power. However, its use would be inappropriate since such a model would only predict intrinsically unacceptable results.

Finally, it is common to misinterpret what the validation of a particular score actually means. This does not refer to the different types of validation methodology to which a model must be subjected, but to the use of the term "validation" by authors who test their models with small populations. Commonly the result is that the model is not validated at all. The validation of a model should be taken to mean that its calibration and discriminating power have been investigated with respect to a certain population and under certain conditions. Firstly, the analysis of its calibration means that the use of the model has been strict, that there has been no artificial increasing of the weight of any variable, and that data have been collected from the entire patient population. Secondly, the examination of its discriminating power (which can be more difficult) requires that losses due to death (however defined) not be left out of any calculations. Daily experience shows that, in the context of complications, it is easy to lose track of the information of a patient with chronic disease who finally dies. If the information-collecting service is imperfect, these patients may not be taken into account. Finally, the size of the sample population is important in the validation of a model. Generally, at least 100 deaths need to be examined.[12] This means that if mortality were 5%, the population sampled would have to include some 2000 patients. Many papers claiming their aim to be the validation of a score have involved much smaller populations.

If validation is performed under appropriate conditions but the model neither calibrates nor discriminates well with respect to the treatment population, then interpretations become very complicated. Two typical scenarios exist: the overestimation or underestimation of risk. If the models used overestimate the true risk, it is usually concluded that our practice is correct. However, the inverse is not always handled in the same way, i.e., if the risk is underestimated, it is often concluded that the models used were badly constructed or

that they were prepared using populations very different from our own. In fact, neither position is entirely correct; the truth lies somewhere in-between.

In conclusion, risk scores are extraordinarily helpful tools whose use is recommended in normal practice. This is true not only when surgery is under consideration, but when any procedure with therapeutic intent (such as the implantation of stents) is contemplated. However, we need to be sure of the reasons for choosing a particular model, and must be aware of the conditions required for its proper use. The limitations of the model and the errors of interpretation associated with it must also be understood.

REFERENCES

1. Eagle KA, Guyton RA, Davidoff R, Edwards FH, Ewy GA, Gardner TJ, et al. ACC/AHA 2004 guideline update for coronary artery bypass graft surgery: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1999 Guidelines for Coronary Artery Bypass Graft Surgery). Circulation. 2004;110:e340-e437.
2. Kennedy JW, Kaiser GC, Fisher LD, Maynard C, Fritz JK, Myers W, et al. Multivariate discriminant analysis of the clinical and angiographic predictors of operative mortality from the Collaborative Study in Coronary Artery Surgery (CASS). J Thorac Cardiovasc Surg. 1980;80:876-87.
3. Kouchoukos NT, Ebert PA, Grover FL, Lindesmith GG. Report of the Ad Hoc Committee on Risk Factors for Coronary Artery Bypass Surgery. Ann Thorac Surg. 1988;45:348-9.
4. Cortina JM, Pérez de la Sota E, Rodríguez E, Molina L, Rufilanchas JJ. Escalas de valoración de riesgo en cirugía coronaria y su utilidad. Rev Esp Cardiol. 1998;51 Suppl 3:8-16.
5. Omar RZ, Ambler G, Royston P, Eliahoo J, Taylor KM. Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. Ann Thorac Surg. 2004; 77:2232-7.
6. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation. 1989;79(6 Pt 2):I3-12.
7. Jones CM, Athanasiou T. Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. Ann Thorac Surg. 2005;79:16-20.
8. Bernstein AD, Parsonnet V. Bedside estimation of risk as an aid for decision-making in cardiac surgery. Ann Thorac Surg. 2000;69:823-8.
9. García Fuster R, Montero JA, Gil O, Hornero F, Cánovas S, Bueno M, et al. Tendencias en cirugía coronaria: cambios en el perfil del paciente quirúrgico. Rev Esp Cardiol. 2005;58:512-22.
10. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg. 1999;16:9-13.
11. Wu Y, Grunkemeier GL, Handy JR Jr. Coronary artery bypass grafting: are risk models developed from on-pump surgery valid for off-pump surgery? J Thorac Cardiovasc Surg. 2004;127:174-8.
12. Harrel FE. Regression models strategies. Berlin: Springer-Verlag; 2001.