Limitations in the inter-observer reliability of EuroSCORE: what should change in EuroSCORE II?

Guillaume Lebreton a, Sylvie Merle c, Jocelyn Inamo b, Jean-Luc Hennequin a, Bruno Sanchez a, Zelie Rilos a, François Roques a,*

a Cardiac Surgery Department, Pierre-Zobda Quitmann University Hospital, Fort-de-France, Martinique, France
b Cardiology Department, Pierre-Zobda Quitmann University Hospital, Fort-de-France, Martinique, France
* Observatoire Régional de la Santé de la Martinique, Fort de France, Martinique, France

Received 20 August 2010; received in revised form 23 February 2011; accepted 25 February 2011; Available online 15 April 2011

Abstract

Objectives: To carry out an in-depth single-centre analysis of the inter-observer reliability of the EuroSCORE (European System for Cardiac Operative Risk Evaluation) to propose changes for the EuroSCORE II. Methods: Data for the EuroSCORE additive and logistic models were prospectively collected by surgeons (computer-assisted calculation) (SurgAE and SurgLE) and perfusionists (on A4 data collection forms; PerfAE) for 1719 consecutive adult heart operations. The performance of the EuroSCORE was first analysed, then inter-observer discrepancies in the score were assessed globally and for any of its 17 risk factors. Results: Hospital mortality was 4.3% (SurgAE and SurgLE: 5.3 and 7.3, respectively). The predictive ability and the calibration of the score were acceptable (area under the receiver operating characteristics curve: 0.75 for SurgAE and 0.753 for SurgLE, \( p = 0.98 \), Hosmer and Lemeshow goodness-of-fit test). Overall inter-observer concordance was satisfactory (Kappa coefficient: 0.71) but SurgAE and PerfAE were different in 26.3% of cases (SurgAE > PerfAE in 18.6%, and PerfAE > SurgAE in 7.7%). Five of the 17 risk factors accounted for most of the variability: left-ventricular ejection fraction, extracardiac arteriopathy, surgery other than isolated coronary artery bypass graft, recent myocardial infarction and pulmonary hypertension (with discrepancies respectively noticed in 7.6%, 5.3%, 5%, 3.9% and 3% of cases). Encoding mismatches for EuroSCORE items have been either assigned to human errors related to interpretation or conflicting information in the charts. Both situations may reflect structural weaknesses of the EuroSCORE. Conclusions: The EuroSCORE is a widely used score, but its predictive power and reliability are declining due to changes in cardiac surgery case mix and outcomes in recent years. The present work highlights the fact that the encoding system in the EuroSCORE still gives room for interpretation. Along with other possible modifications described elsewhere, it is suggested that reliability and predicting ability of the score might be increased by changes in some definitions of risk factors and by the use of numeric values instead of intervals of values.

Keywords: Statistics; Risk analysis/modelling; Surgery; Co-morbidities

1. Introduction

In the 1990s, interest in risk assessment and quality-of-care analysis increased dramatically among health-care providers, insurance companies and physicians involved in cardiac surgery. As a result, there was a need for a simple, user-friendly, objective and predictive overall score. The EuroSCORE (European System for Cardiac Operative Risk Evaluation), produced in its additive version in 1999 [1] and its logistic version in 2003 [2], responded to this quest. This score is now used throughout Europe and also on other continents [3–6].

The EuroSCORE has been applied to different domains of cardiac surgery [7–11]. Recently, it has been used to compare conventional cardiac surgery with alternative procedure [12,13], as, for instance, trans-aortic valve implantation (TAVI). Those studies highlighted an over-estimation of the operative risk by the EuroSCORE. Other studies have reported a loss of predictive power of the EuroSCORE [14–17] due to improvement in skills, to the impact of open reporting of clinical outcome on medical behaviours and quality of practices and to changes in case mix [18]. However, none of those analyses investigated the potential part of subjectivity in the encoding method inside the EuroSCORE system and its possible impact on the score’s predictive power. The analysis of inter-observer differences is a conventional method to highlight subjectivity in reporting systems. This inter-observer study investigates discrepancies in encoding EuroSCORE and its individual risk.
factors. Structural weaknesses of the score will be discussed (objectivity and definition of criteria) to provide information that may be useful for the development of the EuroSCORE II.

2. Materials and methods

Between January 2002 and September 2008, data for the EuroSCORE were collected in our department for 1719 consecutive adult patients. The data were collected independently by surgeons and perfusionists in the preoperative period, according to recent information available in the chart. Additive and logistic models were prospectively collected by the surgeon (computer-assisted calculation; SurgAE and SurgLE) and data for EuroSCORE additive were collected by the perfusionists (on A4 paper forms; PerfAE).

Data were then entered in a local database and exported for analysis on STATA®software.

Risk factors were subjected to univariate analysis (chi-square test for categorical data, Student’s t-test for continuous data). Multivariate analysis was conducted for risk factors that were predictive in univariate analysis ($p < 0.05$). The predictive power of the EuroSCORE was determined by measuring the area under the receiver operating characteristics (AUROC) curve and calibration was carried out with the Hosmer and Lemeshow test (poor calibration if $p < 0.05$). A study of the inter-observer variability of EuroSCORE additive determined by the surgeon (SurgAE) and the perfusionist (PerfAE) was conducted by analysis of the overall agreement (Kappa coefficient $>0.05$) and inter-observer variability was then analysed by item. The matching was analysed for 17 items in 1719 pairs of EA. For categorical risk factors (present or absent), the discordance ratio (number of mismatches/number of cases of occurrence of the risk factor) was measured. In case of inter-observer disagreement, charts were reviewed to specify the origin of discrepancies. Mechanism of the discordance for each items were analysed according to the following classification:

- Type I: Analysis of the charts provides the evidence that the encoding mismatch is related to an error made by one or two of the observers.
- Type II: Analysis of the charts provides the evidence that the encoding mismatch is due to time-related changes in value for an item (i.e., pulmonary systolic pressure, serum creatinine and ejection fraction (EF)).
- Type III: Analysis of the charts provides the evidence that the encoding mismatch is due to differences observed with two different methods of calculation of the value of an item (i.e., EF by echocardiography or angiography).
- Type IV: There is insufficient data in the chart to explain the mismatch.

3. Results

In the study population of 1719 patients, in-hospital mortality (HM) was 4.3%. EuroSCORE additive (SurgAE) was 5.3 and EuroSCORE logistic (SurgLE) was 7.3. The HM/AE ratio was 0.81 and HM/LE ratio was 0.59.
artery hypertension (3.0%). In categorical risk factors, the discordance ratio greatly varies from 0.06% (female gender) to 50.8% (neurological disorder) (Table 3). Among 29,233 pairs of values, 701 single risk factors mismatches were found. They were related to errors by one or (two) of the observers in 398 cases (57.0%). The mismatches were assigned to conflicting information in the charts associated with time related changes in the preoperative period in 141 cases (20.1%) and to discrepancies in the evaluation of an item by different methods of measure in 114 cases (16.3%).

4. Discussion

In the past, it has frequently been observed that the use of a score for the evaluation of practices in cardiac surgery had an impact on how users encode. An overestimation of risk may then occur, leading to a loss of accuracy of the risk model. The most-publicised example was the logistic model proposed in New York State [19] in the early 1990s, which gave rise to a ranking of New York cardiac surgeons (published in the New York Times). Subsequently, a shift in patients’ risk profile was observed for some surgeons, who improved by the way, their ranking. As a consequence, the performance of the model declined, justifying regular revision of the score with changes in the choice and weighting of variables. One of the suspected mechanisms was that those surgeons were inclined to cheat for individual items of the score, which definitions were subjected to observer interpretation (i.e., unstable angina). The EuroSCORE has been widely used to analyse quality of care provided by centres or physicians as a simple institutional audit tool as well as an integral part of a nationwide programme (i.e., Great Britain or France). On other grounds, in recent years, several studies have documented the fact that the predicting strength of the EuroSCORE was declining [14—17]. The ‘overestimation’ of risk has been mostly assigned to improvements in practices during the past decade. The possible implications of structural weaknesses of the score (i.e., definitions, which give a way to interpretations) have never been explored.
Surprisingly, in this study conducted at a centre involved in the EuroSCORE programme, inter-observer differences were observed in as much as 26% of patients. Such variability may be explained by the ‘real-life’ experience purposely used for this study, where observers were completely free of any kind of control. As SurgAE was more often superior to PerfAE than the reverse, one may hypothesise that surgeons attempted to game the scoring system. By reviewing charts, it was only noticed that surgeons scrutinised more deeply the data. In fact, in the few days or weeks preceding an operation, the patient’s condition may not be stable. Thus continuous factors relevant for EuroSCORE (EF, serum creatinine and pulmonary systolic pressure) may greatly vary (sometimes as the simple result of medical preoperative treatment).

Moreover, the measure of a factor may vary, if collected by different methods (echocardiography vs angiography for EF). Those variations (even minimal) will lead to important changes in the score when two measures are located on both sides of a pivotal value (i.e., 200 µm m⁻¹ for serum creatinin, 60 mm of mercury for pulmonary systolic pressure and 30% for EF). In our study, those type II and III discordances explain 36.3% of observed inter-observer discrepancies (Table 3). The use of numerals instead of intervals of values for continuous data of the score may therefore be suggested to avoid this ‘pivot effect’. This would improve the objectivity of the EuroSCORE and decrease its vulnerability to cheating. Table 3 shows that errors made by one or two of the observers (type I discordance) represent 57% of the observed discordances. Those errors cannot be simply assigned to the class of elementary errors or lack of concentration. Thus, definitions may still give room to errors by interpretation. First, some definitions remaining unclear were responsible for encoding mismatches at the beginning of the study. Remarkably, it was not clear to users at that time that the item ‘surgery other than isolated CABG’ should be encoded when the intervention itself was also the subject of dedicated encoding (e.g., surgery of the thoracic aorta or septal rupture after infarction). Oddly, in case of aortic valve replacement, some surgeons encoded surgery for thoracic aorta when an aortic patch was to be used to enlarge the aortic annulus and/or to close the aortotomy. In this study, there was considerable disagreement among observers about the scoring of unstable angina and recent infarction (respectively 4.9% and 10%). Confusion in encoding was probably the result of recent changes in the international definition of myocardial infarction and unstable angina in the new overall concept of acute coronary syndromes [20]. Whether or not those resulting variabilities in scoring have affected the performance of the score itself is questionable. In fact, the predictive power of the score was no more than acceptable (AUROC = 0.705), though this can be explain by the small size of the population (1719 patients) or by a ‘centre effect’. For instance, ethnic origin (African migrants), elsewhere regularly described as an isoleat risk factor for cardiac surgery, may also affect the score performance [21]. In our population, severe distal peripheral vascular diseases are also often observed on completely asymptomatic patients, leading to diverging encoding for this item (Table 3). The impact of wrong encoding for individual EuroSCORE items on the overall predicting ability of the score should increase in case of associated high incidence of the risk factor, high mismatch ratio and high weight of the given factor in the score. For instance, neurological disorder and active endocarditis either requiring a functional assessment by the user (the first) or affected by physician decision making (the latter) varied greatly (Table 3). Their definition might therefore be revisited. At least this work suggests that the EuroSCORE must be integrated in a robust quality assessment programme with a preoperative external control of the encoding process. There is no evidence that most articles highlighting poor performance of the EuroSCORE, particularly those comparing conventional cardiac surgery with alternative techniques, really do.

In conclusion, 10 years after its creation, the EuroSCORE, still remains predictive. However, its use in 2010 in a similar way to 1999 leads to an overestimation of mortality. Modifications in weighting of variable are required to comply with the improvement in outcomes. The score should also take into account changes in the population profile (older patients undergoing heart valve surgery). The logistics environment is also different. Computers and the Internet have now entered the operating room and these new tools can be used to optimise the use and performance of the EuroSCORE II. Thus, the integration of digital data using numerals rather than intervals of values may help to erase the described ‘pivot’ effect. The present study suggests that some definitions should also be revisited.

References

comparison to conventional mitral valve repair or replacement at five years. Thorac Cardiovasc Surg 2006;54:10—4.


