Performance of the original EuroSCORE

Sabrina Siregara,b,*, Rolf H.H. Groenwoldb, Frederiek de Heera, Michiel L. Botsb, Yolanda van der Graafb and Lex A. van Herwerdena

a Department of Cardio-Thoracic Surgery, Heart and Lungs Division, University Medical Center Utrecht, Utrecht, The Netherlands
b Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
* Corresponding author. Department of Cardio-Thoracic Surgery, Division Heart and Lungs, University Medical Center Utrecht, Heidelberglaan 100, E03.511, PO Box 85500, 3508 GA Utrecht, The Netherlands. Tel: +31-88-7556179; fax: +31-88-7555058; e-mail: s.siregar@umcutrecht.nl (S. Siregar).

Received 24 June 2011; received in revised form 22 November 2011; accepted 24 November 2011

Summary

The European system for cardiac operative risk evaluation (EuroSCORE) is a commonly used risk score for operative mortality following cardiac surgery. We aimed to conduct a systematic review of the performance of the additive and logistic EuroSCORE. A literature search resulted in 67 articles. Studies applying the EuroSCORE on patients undergoing cardiac surgery and which reported early mortality were included. Weighted meta-regression showed that the EuroSCORE overestimated mortality. However, this performance depended on the risk profile of patients: in high-risk patients, the additive model actually underestimated mortality. Discriminative performance was good. Given the poor predictive performance, the EuroSCORE may not be suitable as a tool for patient selection nor for benchmarking.

Keywords: Prognostic factors • CABG • Valves • Outcome • Review

INTRODUCTION

In recent decades, multiple risk scores have been developed to estimate the outcome of cardiac surgery [1]. A commonly applied risk score for operative mortality after cardiac surgery is the European system for cardiac operative risk evaluation (EuroSCORE) [2]. It was developed by a European steering group in 1999 with information on approximately 15 000 consecutive adult patients undergoing cardiac surgery under cardiopulmonary bypass.

In 1995, a database was formed with information from 128 centres in eight European countries. The database was divided into a derivation subset (n = 13 302) and a validation subset (n = 1497). Statistical analyses identified 17 characteristics that were included in the model. These variables were then each given a weight based on the logistic regression beta-coefficients to form an additive risk score for operative mortality. The final model was tested on both the construction and the validation subsets and showed a satisfactory performance [2]. However, as stated by the authors, the true test lies in the widespread use of such a risk score.

The purpose of the EuroSCORE is to help in the assessment of the quality of cardiac surgical care [2]. This can be done by comparing observed mortality rates against the expected rates between care providers, the so-called benchmarking. For this purpose, however, the presence of other factors that might influence outcome should always be kept in mind, for example, post-operative care.

The quality of predictions of a widely used risk score has major consequences. Poor performance leads to inadequate prediction and potentially to invalid benchmarking. An appropriate evaluation of the performance should include all available evidence. So far, one systematic review, limited to six articles, has been written on this topic [3]. It does not fully comprise the large amount of literature on this subject.

The new EuroSCORE is published simultaneously in this issue [4]. This is a crucial moment to understand which improvements have and should have been made to the model. Therefore, the objective of this study was to systematically review all available literature on the performance of the additive and logistic EuroSCORE.

MATERIALS AND METHODS

Search strategy and study selection

On 6 June 2010, a search was conducted in PubMed. The search terms ‘EuroSCORE’ and ‘Euro SCORE’ were used in the fields of title and abstract. Based on title and abstract, all studies applying the additive or logistic EuroSCORE on patients undergoing cardiac surgery were included for further evaluation.

The following exclusion criteria were applied: articles written in languages other than English, studies conducted in the original EuroSCORE database and editorial comments or letters. Articles with very specific domains such as endocarditis, thoracic aortic surgery and prosthesis dysfunction were excluded as well.

Remaining studies focused on coronary artery bypass grafting (CABG), valvular surgery and combined surgery. Both prospective and retrospective studies were included. Only studies that
reported early mortality were included in the final selection. Early mortality was defined as either 30-day mortality, in-hospital mortality or operative mortality (30-day and in-hospital mortality). If no expected and observed mortality could be extracted from the text or could be calculated from the presented data, articles were excluded. In case multiple articles reported on the same dataset, the first published article was selected.

The predictive performance of the EuroSCORE was quantified by means of calibration and discrimination.

Calibration

Calibration refers to the ability of a test or a model to estimate the probability of the occurrence of the outcome, in this case mortality [5]. A well-calibrated model for mortality is one with a high agreement between the actual (observed) number of deaths and the predicted (expected) number of deaths. In this review, we focused on aggregate data instead of individual patient data, i.e. the mean observed mortality and the mean expected mortality of included studies. The calibration of the EuroSCORE was assessed by the observed: expected ratio (O:E ratio) of mortality. This ratio was obtained by dividing the observed mortality with the expected mortality within a population. Ideally, this ratio equals one, since in that case the observed mortality equals the expected mortality and hence the predictive model is optimally calibrated. A value below one corresponds to overestimation of mortality and a value above one to underestimation of mortality. The confidence interval (CI) of the ratio was then estimated using the method by Breslow and Day [6]. Whenever the expected mortality or mean EuroSCORE was not explicitly mentioned in an article, the mean additive EuroSCORE was estimated based on reported distributions of pre-operative patient characteristics if possible. Expected mortality was then calculated by incorporating the mean values of the patient characteristics in the additive EuroSCORE model.

The calibration of the EuroSCORE was also evaluated for different risk groups. For this, articles were used that stratified patients by their pre-operative risk (in EuroSCORE), provided that observed mortality, expected mortality and size of each stratum were reported. The O:E ratio calculated from each stratum of patients represented one measurement in the analyses.

Discrimination

Discrimination refers to the ability of the EuroSCORE to differentiate between post-operative survivors and non-survivors [5]. This was quantified using the reported areas under the receiver operating characteristic curve or c-statistics. A c-statistic of 0.5 indicates no ability of the model to discriminate, and a c-statistic of 1.0 indicates a perfect ability to discriminate.

Surgical categories

Studies or cohorts described within studies were divided according to the performed procedure. The following categories were identified: cardiac surgery, isolated CABG, isolated valve and mixed CABG and valve. The cardiac surgery category contained studies that included all forms of cardiac surgery. Some of these articles excluded specific procedures such as thoracic aortic surgery, off pump surgery and surgery for congenital anomalies. The mixed CABG and valve category comprised all studies that could not be allocated to the surgical categories cardiac surgery, isolated CABG or isolated valve.

Analyses

All analyses were conducted using PASW Statistics 17.0 [7] and R for Windows [8]. To evaluate the calibration of the EuroSCORE, we assessed the relation between predicted mortality and the O:E ratio using meta-regression analysis. Additionally, the relation between year of surgery (defined as the median of the years of surgery included in a study) and O:E ratio was assessed, also by means of meta-regression analysis. All meta-regression analyses were performed using a random effects model, in which the O:E ratio for each study was weighted by the inverse of the variance of the O:E ratio. In practice, this means that larger studies tend to get more weight. Since the O:E ratio was not normally distributed, we log-transformed this ratio, after which it was normally distributed. Furthermore, studies that presented O:E ratios stratified by risk score (i.e. stratified by expected risk) were analysed separately using a mixed random effects model, thus accounting for the dependency of the stratified observations within studies.

To evaluate the discriminative performance of the EuroSCORE, the c-statistic of each study was weighted according to study size. The variance of the c-statistic was not routinely reported and could not be calculated from the reported aggregate data.

Sensitivity analyses were performed to evaluate whether the findings were mainly determined by the largest studies. This was done by excluding the studies with the largest weight and repeating all analyses. These results were then compared with the original results.

RESULTS

Overview of available literature

The search resulted in 686 articles (Fig. 1). After applying the inclusion and exclusion criteria, 102 articles remained. Among these studies, seven were not available in full text [9–15]. Another 10 studies had a domain that was too specific [16–25], and 11 studies did not report an observed or an expected mortality rate (nor could these be calculated based on what was reported) [26–36]. Seven articles gained data from a data set used for a previously published article.

The final evaluation thus included 67 articles [1, 37–102], which were based on surgery performed from 1992 to 2009 on 462 243 patients. The studies varied from large multi-centre studies conducted to assess the overall performance of the EuroSCORE to single-centre studies using the EuroSCORE for internal quality control.

Studies applied either the additive or the logistic EuroSCORE, or both. The mean expected EuroSCORE was not reported in nine articles and had to be estimated from the reported data.

Calibration of the EuroSCORE

In Table 1, the results of data extraction and analysis are shown. Most studies applied the additive EuroSCORE and the general
domain of cardiac surgery (21 articles). The mean observed mortality was 3.9 and 3.1% for studies applying the additive (53 articles) and logistic EuroSCORE (47 articles), respectively. The mean expected mortality was approximately twice as high and resulted in O:E ratios of 0.47 and 0.43, respectively. The overestimation is more evident for the logistic score than for the additive score. Although calibration of the EuroSCORE varied much between the surgical categories, all the means of the O:E ratios were below 1.0. This is clearly illustrated in Figs S1 and S2 of the Supplementary material.

Overall, O:E ratios decreased with increasing score \[ \text{O:E ratio} = e^{-0.067 \times \text{additive EuroSCORE} - 0.199} \] (95% CI of coefficient \((-0.090; -0.044)) and \[ \text{O:E ratio} = e^{-0.041 \times \text{logistic EuroSCORE} - 0.568} \] (95% CI of coefficient \((-0.072; -0.010))]. This means that overestimation is stronger in high-risk patients. When the analyses were performed with studies stratifying in risk groups \([39, 40, 44, 45, 47, 48, 50, 57, 59, 64-68, 71, 73, 75, 76, 85, 90, 95-100] \), the O:E ratios of the additive model increased with increasing score (Fig. 2). An O:E ratio of more than 1.0 was seen in scores above 15.2. This means that the degree of overestimation is actually less in high-risk patients and that even an underestimation of mortality is found above an additive score of 15.2. The logistic EuroSCORE demonstrated O:E ratios between 0.4 and 0.5, with decreased ratios in high-risk patients, which is illustrated in Fig. 3. Sensitivity analysis showed similar results when the largest studies were excluded. For the separate surgical categories, the direction of the regression lines corresponded with those in Figs 2 and 3, although CIs were wider due to the fewer studies available.

The calibration of the additive EuroSCORE throughout the years is depicted in Fig. 4. The O:E ratio has been below 1.0 (i.e. overestimation of mortality) since 1994 and is slowly increasing \[ \text{O:E ratio} = e^{0.049 \times \text{year of surgery} - 1994} - 0.900 \] (95% CI of coefficient \((0.030; 0.067))]. This means that the degree of overestimation has declined. When the largest study \([72] \) was excluded from the analyses, a similar increasing trend was found \[ \text{O:E ratio} = e^{0.041 \times \text{year of surgery} - 1994} - 0.582 \] (95% CI of coefficient \((-0.031; 0.051))].

---

**Table 1: Observed and expected mortality rates for the EuroSCORE by surgical category**

<table>
<thead>
<tr>
<th>Surgical Category</th>
<th>No. of patients</th>
<th>No. of articles</th>
<th>Observed mortality (%)</th>
<th>Expected mortality (%)</th>
<th>O:E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean^b</td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td><strong>Additive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All studies</td>
<td>373 531</td>
<td>53</td>
<td>3.9</td>
<td>0.8</td>
<td>10.6</td>
</tr>
<tr>
<td>Cardiac surgery</td>
<td>268 928</td>
<td>21</td>
<td>4.1</td>
<td>1.1</td>
<td>5.9</td>
</tr>
<tr>
<td>Isolated CABG</td>
<td>75 972</td>
<td>18</td>
<td>2.0</td>
<td>0.8</td>
<td>4.9</td>
</tr>
<tr>
<td>Isolated valve</td>
<td>8633</td>
<td>7</td>
<td>2.9</td>
<td>2.2</td>
<td>4.8</td>
</tr>
<tr>
<td>CABG and valve</td>
<td>26 502</td>
<td>9</td>
<td>4.5</td>
<td>3.5</td>
<td>10.6</td>
</tr>
<tr>
<td><strong>Logistic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All studies</td>
<td>193 814</td>
<td>44</td>
<td>3.1</td>
<td>0.6</td>
<td>13.9</td>
</tr>
<tr>
<td>Cardiac surgery</td>
<td>80 613</td>
<td>11</td>
<td>3.5</td>
<td>2.5</td>
<td>7.5</td>
</tr>
<tr>
<td>Isolated CABG</td>
<td>96 062</td>
<td>12</td>
<td>2.3</td>
<td>0.8</td>
<td>5.0</td>
</tr>
<tr>
<td>Isolated valve</td>
<td>16 703</td>
<td>11</td>
<td>2.8</td>
<td>0.6</td>
<td>3.9</td>
</tr>
<tr>
<td>CABG and valve</td>
<td>12 773</td>
<td>12</td>
<td>5.7</td>
<td>3.52</td>
<td>13.9</td>
</tr>
</tbody>
</table>

^aStudies excluded with only high risk or high age inclusion.

^bMeans are weighted (see Materials and methods). Bold values are used merely to emphasize the means.
Most results for the logistic EuroSCORE were non-significant because of the small number of studies. A small significant increase in the calibration over the years was found in the isolated CABG category $[O:E \text{ ratio } = e^{0.047 \cdot (\text{year of surgery} - 1997) - 0.947}, 95\% \text{ CI of coefficient } (0.032; 0.092)]$. One circle represents one risk stratum and the size indicates its weight.

The expected mortality over time was also analysed with regression analysis. The results were significant and showed a decrease in the reported expected mortality over the years, which is depicted in Fig. 5. Studies using the additive model showed the same decrease in expected mortality over time [additive EuroSCORE $= e^{0.002 \cdot \text{logistic EuroSCORE} - 0.897}, 95\% \text{ CI of coefficient } (-0.003; -0.001)]$. One circle represents one risk stratum and the size indicates its weight.
Discrimination of the EuroSCORE

The discrimination ability of the EuroSCORE was good, with average $c$-statistics between 0.7 and 0.8 in all surgical categories except for the logistic EuroSCORE in the isolated valve category (Table 2). Over the years, the mean $c$-statistic of all studies remained 0.8 ($c$-statistic of additive EuroSCORE $= 0.002 \times \text{(year of surgery − 1994)} + 0.772, 95\% \text{ CI of coefficient } (-0.0004; 0.004)$ and $c$-statistic of logistic EuroSCORE $= -0.005 \times \text{(year of surgery − 1997)} + 0.803, 95\% \text{ CI of coefficient } (-0.010; 0.0004)$).

**DISCUSSION**

**Principle findings**

This systematic review shows that both the additive and the logistic EuroSCORE overestimate mortality in all surgical categories. The degree of overestimation depends on the pre-operative risk of the patients. Furthermore, the overestimation is more prominent by the logistic than by the additive EuroSCORE. Previous reports on the performance of the EuroSCORE similarly concluded that the EuroSCORE overestimated mortality [3, 45]. In high-risk patients, an overestimation of mortality by the additive EuroSCORE has previously been shown as well [3, 46, 62, 66, 100]. It was suggested that this phenomenon is inherent to the additive model [62] and thus can be resolved by using a logistic method [103]. This systematic review shows that the logistic model overestimates mortality in all risk groups. Indeed, calibration of the logistic model appears to be less dependent on the pre-operative risk of the patient, which means that it is more stable across risk groups. For these reasons, it is advisable that the logistic model is used whenever possible.

**Trends in past years**

A common explanation for the described overestimation is the hypothesis that the EuroSCORE is an outdated risk score. Therefore, a good calibration at the start and deterioration with progressing years were expected in the analyses. The EuroSCORE was constructed using data from patients operated in 1995. Changes in indication for cardiac surgery and the increasing role of percutaneous intervention might have had an effect on patient characteristics. Technological improvements of pre-, peri- and post-operatively used equipment have likely reduced the risk of mortality. These changes could partly have accounted for the poor calibration in current practice [3]. However, our analysis demonstrated an opposite trend. The additive and logistic EuroSCORE showed $O/E$ ratios below 1.0 from the beginning and a slow increase in $O/E$ over the years. In other words, overestimation was present from the beginning (Fig. 4), and deterioration of the score over time could not be demonstrated despite the 16 years that have passed.

Another remarkable trend was found when studies were observed over time. Our results indicate a decrease in expected risk when all studies were pooled. Subgroup analyses for all surgical categories showed similar trends, although results were not significant. Previous reports from other large cohorts are ambiguous regarding this matter. The database of the Society of Thoracic Surgeons in the USA demonstrated that isolated CABG patients are currently older and sicker than before [104]. The Society of Cardiothoracic Surgeons in Great Britain and Ireland reported an increase in patients with high risk in all surgical categories [105]. However, other large cohorts in Europe report no or a minimal increase in patient risk over time. In the Danish Heart Register, a marginal increase of 0.02 is detected in the mean additive EuroSCORE from 2006 to 2010 (from 5.16 to 5.18 for isolated CABG and from 7.03 to 7.05 for valvular surgery) (The Danish Heart Register: [http://www.dhreg.dk/](http://www.dhreg.dk/)). The Netherlands Association for Cardio-Thoracic Surgery reports no rise in the logistic EuroSCORE from 2007 to 2009 (both 4.9% for isolated CABG and from 8.0 to 7.8% for aortic valve replacement) (The Netherlands Association for Cardio-Thoracic Surgery: [http://www.nvtnet.nl/](http://www.nvtnet.nl/)).

**Valvular surgery**

The EuroSCORE derivation data set consisted of 63.6% isolated CABG and 29.8% valvular operations. It has previously been discussed that the type of procedure affects the outcome and that the influence of a risk factor differs across the types of procedures [105]. We found that the EuroSCORE performed worst in
the isolated valve category regarding both calibration and discrimination. In a comment by Nashef [106], a less favourable discrimination in valvular surgery was explained by exclusion of certain items of the EuroSCORE that allow the model to discriminate. However, studies that do not exclude EuroSCORE items also found low c-statistic of 0.69 [57, 75]. Despite these examples of poor discrimination, the mean c-statistic for both the EuroSCORE models remains above 0.7 (Table 2) and thus sufficient.

Some authors advocated the use of separate risk stratification models for valvular surgery [93]. However, the strength of the EuroSCORE is its general applicability to various kinds of cardiac surgery. Extra attention to risk factors related to valvular surgery should make it possible to keep one model with acceptable performance for cardiac surgery.

### Purpose of the EuroSCORE

A model that is constructed for patient selection must meet other requirements than a risk model that is constructed for benchmarking. The first type of model should be simple enough for clinical use, and both calibration and discrimination are important. For the latter, the main concern is a scrupulous calibration in all risk groups, so that fair comparisons between providers can be made [107].

The reported aim of the EuroSCORE was to aid quality assessment in surgical care. For this purpose, the demonstrated good discrimination alone is not adequate and calibration is of higher importance. Unfortunately, the weakness of the EuroSCORE model is its calibration. The overestimation cannot simply be resolved by multiplying the expected risk with a certain factor (e.g. in this case multiplying the expected mortality by a factor 0.5). Because the calibration differs across risk groups, the correction factor would have to be different for every risk group. Consequently, the model may not be accurate enough for benchmarking. It remains debatable whether the calibration difference across risk groups in the logistic EuroSCORE is clinically relevant for the sole purpose of patient selection. In this issue, EuroSCORE II is presented [4]. The goal of improving both calibration and discrimination fits the initial purpose of the score well.

### Improving the model: EuroSCORE II

It is evident that improvement was required and has indeed been made with the updated model, EuroSCORE II [4]. First, the score was developed using more patients from a diverse collection of countries all over the world. This increases generalizability of the model. Furthermore, the authors omitted the additive score. The problem regarding the calibration of the additive score, as discussed previously in this paper, is therefore no longer an issue. Thirdly, in order to enhance the performance of the model in valvular (and other concomitant) surgery, the number of major cardiac procedures performed is incorporated in the score calculation. Although other specific risk factors for valvular surgery were not added to EuroSCORE II, this modification is likely to improve the performance of the model in valvular surgery. At the same time, the convenience of having one model for different types of cardiac surgery is maintained.

Considering the calibration problems discussed in this article, the largest improvement is probably made with the total recalibration of the model. The intercept as well as all coefficients have been updated. This caused the calibration of the model to improve dramatically: the observed: expected ratio in the validation set is 0.94. Whether this can be reproduced in other data sets remains to be seen; yet, this result is promising. In addition, several risk factors have been added or altered in some way (e.g. left ventricular function is now divided into four categories). Although the c-statistic of the former EuroSCORE was satisfactory, this is likely to further improve the discrimination of the model. Taking these changes together, the EuroSCORE has been improved in many ways and we expect the performance of EuroSCORE II to be superior to the former model.

### Limitations and strengths

Potential limitations of this review include publication bias, quality of the reviewed studies and ecological bias. Publication bias is a potential problem in every meta-analysis, but particularly in meta-analyses on observational studies when compared with those on randomized trials [108, 109]. Since statistical significant results are more likely to be published, the reviewed papers might not reflect the actual situation. Recently, Nashef [106] commented on the performance of the EuroSCORE in valvular operations and also stated that publication bias might explain the overestimation. In this review, however, funnel plots of the O:E ratios were not suggestive for publication bias. However, as can be seen from the comparison of trends between our results and those from other cohorts, it remains unclear whether the reviewed papers are representative of all cardiac surgery. Therefore, publication bias should always be considered in the interpretation of results.

The validity of this review is directly affected by the methodological quality of the included studies. According to guidelines on the conduct of prognostic studies, this feature is related to the representativeness of the study populations, loss to follow-up and measurement of prognostic factors and outcomes. Referring to the former, it is impossible to evaluate whether the studied populations are representative for entire cardiac surgery population. As to the second point, it is unlikely that loss to follow-up materially affected our results. The reason for this is that the studied outcome was early post-operative mortality, which means that follow-up time is short. Lastly, the measurement of predictors and outcome was inconsistent across articles. Some studies extracted data from existing databases that used other definitions than those in the EuroSCORE. Outcome differed from in-hospital mortality to 30-day mortality and classic operative mortality. The impact of these differences in methodological quality is difficult to evaluate.

In this review including meta-regression analysis, aggregate data were used, i.e. no individual patient data. This may have led to ecological bias: the effect observed in the aggregate data is biased and would not be present if individual patient data were analysed. To evaluate the calibration of the EuroSCORE across risk groups, an additional analysis was therefore performed using O:E ratios derived from each risk group within a study. The advantage of this analysis is that smaller numbers of patients are aggregated, which likely reduces possible ecological bias. Indeed, for the additive EuroSCORE, the overestimation of mortality risk appeared constant when considering aggregate data.
only, whereas this overestimation appeared to be stronger in high-risk patients when considering stratified data. For the logistic EuroSCORE, such discrepancy was not observed. Reported data were usually not stratified in time periods or years. Therefore, similar additional analyses for trends in time could not be performed with time-stratified data. The phenomenon of ecological bias is a common pitfall in the analysis of aggregate data, and its occurrence should always be considered with caution.

One of the strengths of this review is its comprehensiveness. In this study, 67 articles were used to evaluate the performance of the EuroSCORE. Hypotheses and results of individual studies could therefore be reliably confirmed or rejected. The only previous review on the EuroSCORE included six studies.

Another advantage of this study is that extracted data were divided according to the surgical category. This gave a clear view on the poor performance of the EuroSCORE in different types of procedures and answered the question whether a separate score is needed for valvular surgery.

Finally, all analyses performed in this review were weighted according to a random effects model. Consequently, data from large studies have more impact on the results than data from small studies. This made it possible to use all available studies, while limiting the influence of small studies.

CONCLUSIONS

This comprehensive review shows that both the additive and the logistic EuroSCORE do not adequately predict operative mortality following cardiac surgery. The discrepancy between the expected and observed mortality differs across risk groups. Therefore, the EuroSCORE may not be suitable as a tool for patient selection and benchmarking of healthcare providers.

SUPPLEMENTARY MATERIAL

Supplementary material is available at EJCTS online.

Conflict of interest: none declared.

REFERENCES


