Our reference: HLC 1581 P-authorquery-v9

# AUTHOR QUERY FORM

| | | |
|---|---|---|
| ELSEVIER | **Journal:** HLC<br>**Article Number:** 1581 | **Please e-mail or fax your responses and any corrections to:**<br>**E-mail:** corrections.esch@elsevier.thomsondigital.com<br>**Fax:** +353 6170 9272 |

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. Note: if you opt to annotate the file with software other than Adobe Reader then please also highlight the appropriate place in the PDF file. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

Any queries or remarks that have arisen during the processing of your manuscript are listed below and highlighted by flags in the proof. Click on the 'Q' link to go to the location in the proof.

| **Location in article** | **Query / Remark: click on the Q link to go**<br>**Please insert your reply or correction at the corresponding line in the proof** |
|---|---|
| Q1 | Please check suggested running title for correctness. |
| Q2 | Please confirm that given names and surnames have been identified correctly. |
| Q3 | Please check affiliations "a" and "b" for correctness. |

Please check this box or indicate your approval if you have no corrections to make to the PDF file □

Thank you for your assistance.

Heart, Lung and Circulation (2014) **xx**, 1–10
1443-9506/04/$36.00
http://dx.doi.org/10.1016/j.hlc.2014.04.005

ORIGINAL ARTICLE

# Performance of EuroSCORE II in Hungary: A Single-centre Validation Study

**György Koszta, MD [a*], Gábor Sira, MD [a], Katalin Szatmári, MD [a], Eszter Farkas, MD [a], Tamás Szerafin, MD, PhD [b], Béla Fülesdi, MS, PhD, DSci [a]**

[a]University of Debrecen, Medical and Health Science Centre, Department of Anaesthesia and Intensive Care, H-4032, Nagyerdei krt 98, Debrecen, Hungary
[b]Department of Cardiac Surgery, H-4032, Nagyerdei krt 98, Debrecen, Hungary

*Received 11 January 2014; received in revised form 12 March 2014; accepted 4 April 2014; online published-ahead-of-print xxx*

**Background** The efficacy of the updated cardiac surgical risk stratification system, EuroSCORE II, needs widespread assessment in the cardiac surgical centres where it is intended to be used. The present paper is a single-centre validation study carried out in Hungary.

**Methods** An adult cardiac surgical cohort of 2287 patients was investigated. The general levels of performance of the logistic EuroSCORE and that of EuroSCORE II were compared using the Hosmer-Lemeshow test, ROC analysis and calculation of the Brier score. The calibrations were visualised by smoothed curves derived with the help of local polynomial regression. The efficacy of EuroSCORE II was analysed in different operation types and urgency subgroups.

**Results** The old EuroSCORE over-estimated the risk (O:E ratio: 0.66, HL test, p<0.01), while EuroSCORE II slightly under-predicted mortality (O:E ratio:1.19, HL test, p=0.0084). Comparing the ROC AUCs, we did not find a significant difference between the accuracy of the old and new versions of EuroSCORE (0.8017, 95% CI:0.7596-0.8438 vs. 0.8177 95% CI: 0.7786-0.8569). EuroSCORE II performed well among CABG patients (O:E ratio: 0.75, HL test, p=0.5789) and in those who underwent elective surgery (O:E ratio: 1.1, HL test, p=0.1396), but failed in the emergency (O:E ratio: 1.71, HL test, p=0.0055) and salvage (O:E ratio:1.36, HL test, p=0.0245) categories.

**Conclusions** EuroSCORE II proved to be more suitable for cardiac surgical risk prediction compared with its previous version, but its reliability can be questioned among patients who need emergency and salvage surgery, as well as in the case of combined operations.

**Keywords** Risk model • Cardiac surgery • EuroSCORE • Mortality • Risk stratification • Validation

## Introduction

The recently published EuroSCORE II [1] refreshed our knowledge of adult cardiac surgical risk and gave us an updated tool for everyday practice. The primary aim of risk stratification is to provide information about the likely outcome for both the patient and the clinicians. The improvement of cardiac surgical care is also based on continuous quality control, in which the expected and the observed outcomes are compared.

*Corresponding author at: University of Debrecen, Medical and Health Science Centre, Department of Anaesthesia and Intensive Care, Nagyerdei krt. 98, H-4012, Debrecen, Hungary. Tel.:/fax: +36 52 411 717/54347, Email: kosztagy@gmail.com

A valid risk stratification system will be of paramount importance in clinical decision making when choosing between cardiac operations and catheter-based interventions, which are now available offering a palliative, but lower-risk solution for an increasing number of cardiac conditions. To confirm the practical usefulness of the new EuroSCORE several external validation studies are needed in different countries.

Our institution participated in the data collection process for the recalibration of EuroSCORE, but neither its present nor its old version [2] has been validated in Hungary to date. The aim of the present study is to examine the calibration and accuracy of EuroSCORE II on a Hungarian adult cardiac surgical population.

We suspected that the performance of EuroSCORE in Hungary might be different from that of other regions in Europe. There are several reasons behind this assumption: (1) different genetic background of the population, (2) socio-cultural aspects, (3) different economic resources of the health care system. These factors are not or poorly represented in the EuroSCORE II. risk model, because the majority of the patients enrolled into the developmental database came from the Western European region, that basically differs from the Eastern and Middle Europe as well as from Asia and Australia in the above-mentioned aspects.

There are certain published data that indirectly suggest strong a genetic influence behind the risk factors of cardiac disease in the Hungarian population. Farsang et al [3] reported increased incidence of cardio-metabolic syndrome in the Central European population compared with other regions of Europe. Beyond the inherited factors the explanations for the unfavourable risk profile in Central-Europe can be life-style (low level of physical exercise and high amount of saturated fat in the diet).

On the bases of the WHO on-line database (*http://www.who.int/countries/en/*) one can explore the differences among the countries which contributed data to the new EuroSCORE. The expenditure on health per capita is around 50% compared with the Western European countries but the hazardous effect the more frequent smoking places on health care is disproportionally higher. The probability of dying between the 15th and the 60th years of life is more than the double (208/1000) for males in Hungary compared with the Western European data (United Kingdom: 91/1000; Austria: 94/1000; Germany: 96/1000; France: 113/1000). These indices in other Central European countries are the following: the Czech Republic: 132/1000; Slovakia: 170/1000; Poland: 191/1000; Romania: 209/1000; the Ukraine: 310/1000. For comparison the same ratio is 80/1000 in Australia.

The present publication aims at exploring the performance of EuroSCORE II independently of these three non-specified determinants.

In addition to describing EuroSCORE II's general performance, we also aimed to explore its efficacy in the different cardiac surgical groups, as well as in the urgency categories.

## Patients and Methods

The recruitment of the validation cohort started on 1st November 2010 and ended on 31st January 2013 in a single cardiac surgical centre, shortly after the data collection for EuroSCORE II ended. All the patients who underwent major cardiac surgical procedures (CABG, AVR, MVR, mitral valve repair, ascending aorta replacement or repair, atrial septal defect closure, atrial myxoma excision, or a combination of these) were enrolled and followed up to the 30th postoperative day. The same risk predictors were collected as had been provided for the developmental EuroSCORE II database, but none of the patients was included in both datasets. All the patients enrolled into this validation study signed an informed consent form in which they agreed with the use of the data collected for their disease, their treatment and outcomes for scientific and publication purposes. The study was approved by the local ethical committee. The only outcome parameter was in hospital mortality within this period.

For the calculation of the logistic EuroSCORE and the EuroSCORE II p-values, the online tools were used that can be found on the website: www.euroscore.org.

The distribution of the calculated risks was depicted on a Logistic EuroSCORE-EuroSCORE II-scatter plot with different symbols for the survivors and non-survivors.

The basic overall performance parameter was the observed to expected mortality ratio (O:E ratio). The practical meaning of the individual logistic EuroSCORE and EuroSCORE II p-value is the probability of death within 30 days following the operation. The expected mortality was calculated by averaging out these probabilities [4].

Calibrations of the scores were evaluated using the Hosmer-Lemeshow test. The expected mortalities in the deciles of the predicted risk were calculated on the basis of both models, similar to above, by averaging the individual p-values and comparing them with the observed mortality in each decile. The difference between the observed and the expected mortality was considered to be statistically significant if the HL-test result was $<0.05$ [4]. In order to demonstrate the goodness-of-fit visually, calibration curves were created by using a smoothing method. These curves are the results of a local polynomial regression where Epanechnikov kernel function was used with a bandwidth of 0.05 [5].

The accuracy or discriminative power of the risk stratification models were analysed by using the receiver operation characteristics (ROC) method. The area under the ROC curves and their 95% confidence intervals were calculated and compared [4].

As another general measure of accuracy the Brier score of each individual outcome prediction was calculated according to the following formulas:

Brier score (BS) = $(p-1)^2$ if the patient died and $(p-0)^2$ if the patient survived, where p is the probability of mortality within 30 days following the surgery, predicted by either the logistic EuroSCORE or EuroSCORE II [4]. The reported BS values are the means of these individual Brier scores. If we know the outcome, the Brier score is zero when the prediction is perfect and – analogous to tossing a coin – it is 0.25 or higher when the prediction is uninformative or even misleading. The frequency of patients with individual BS ≥0.25 is also reported.

Cardiac surgical subgroups were created according to the weight of intervention option (isolated coronary, single non-coronary, two procedures, three procedures) and on the basis of the procedural urgency (elective, urgent, emergent, salvage) and the same calibration and accuracy parameters were calculated, but only in connection with EuroSCORE II.

If the size of a certain group was too small, the HL-test was modified: the individual risks were sorted into five of three groups rather than into deciles, in order to gain appropriate statistical power in each subgroup. The results of the modified HL-tests were indicated with *, if five and **, if three groups were used.

The continuous numerical data were reported as means along with their standard deviation, while the categorical data were reported as frequencies and their relevant percentages throughout the text. The results of the statistical tests were considered to be significant if $p<0.05$.

The collection and processing of the data were performed with the help of the STATA 10 statistical package (STATA Corp., Texas, USA).

# Results

## General Description of the Cohort, Frequency of Risk Predictors and Distribution of the Risk

The frequency of the risk predictors are summarised in Table 1. A total of 2287 patients were included in the analysis, 1491 males and 796 females. Females were significantly older than males (61.7±10.1 years vs. 64.4±9.8 years, $p<0.01$) and had a higher risk according to both scores. Logistic EuroSCORE probabilities were 0.0727±0.1037 for males vs. 0.0982±0.1380 for females, $p<0.001$. EuroSCORE II probabilities were 0.0397 ±0.0619 for males vs. 0.0553±0.0901 for females, $p<0.001$.

Table 2 lists the cardiac surgical procedures and their frequencies in the validation cohort. Single coronary operations were performed in 1038 cases (45.4%), the majority of which were on-pump procedures. Aortic valve replacement (AVR) was the most frequent non-coronary intervention.

The expected mortality of 8.18% (95% CI: 7.7-8.6%) as calculated by the logistic EuroSCORE, was significantly higher than the 4.5% (95% CI: 4.2% -4.8%) mortality predicted by EuroSCORE II, $p<0.01$. 123 persons (5.4%) died within 30 days following cardiac surgery.

The scatter plot in Fig. 1 depicts a comparison of the new and old scores. The logistic EuroSCORE indicated higher risks in the majority of the cases. The EuroSCORE II p-values were higher in only 244 Individuals, of whom 109 died.

## Calibration of the Logistic EuroSCORE and EuroSCORE II

Table 3 summarises the results of the Hosmer-Lemeshow test of the old logistic EuroSCORE. The logistic EuroSCORE

**Table 1 Basic characteristics of the cohort.**

| RISK PREDICTOR | |
|---|---|
| age | |
| years (SD) | 62.7 (10.1) |
| gender | |
| male, n (%) | 1491 (65.2) |
| female, n (%) | 796 (34.8) |
| creatinine clearence | |
| >85, n (%) | 1143 (50) |
| 85-50, n (%) | 862 (37.7) |
| <50, n (%) | 254 (11.1) |
| on HD, n (%) | 28 (1.2) |
| extracardiac arteriopathy | |
| n (%) | 769 (33.6) |
| poor mobility | |
| n (%) | 64 (2.8) |
| previous cardiac surgery | |
| n (%) | 121 (5.3) |
| chronic lung disease | |
| n (%) | 373 (16.5) |
| active endocarditis | |
| n (%) | 74 (3.2) |
| critical preoperative state | |
| n (%) | 81 (3.5) |
| diabetes on insulin | |
| n (%) | 304 (13.3) |
| NYHA grade | |
| I, n (%) | 210 (9.2) |
| II, n (%) | 911 (39.8) |
| III, n (%) | 1079 (47.2) |
| IV, n (%) | 87 (3.8) |
| angina at rest | |
| n (%) | 264 (11.5) |
| ejection fraction | |
| >50%, n (%) | 1300 (56.8) |
| 31-50%, n (%) | 879 (38.4) |
| 21-30%, n(% | 100 (4.4) |
| <20%, n (%) | 8 (0.4) |
| MI within 90 days | |
| n (%) | 301 (13.2) |
| pulmonary hypertension | |
| <30 mmHg, n, (%) | 1423 (62.2) |
| 31-55 mmHg, n (%) | 670 (29.3) |
| >55 mmHg, n (%) | 194 (8.5) |
| urgency | |
| elective, n (%) | 1694 (74.1) |
| urgent, n (%) | 488 (21.3) |
| emergent, n (%) | 77 (3.4) |
| salvage, n (%) | 28 (1.2) |
| surgery on the thoracic aorta | |
| n (%) | 115 (5) |
| postinfarct septum rupture | |
| n (%) | 18 (0.8) |

Please cite this article in press as: Koszta G, et al. Performance of EuroSCORE II in Hungary: A Single-centre Validation Study. Heart, Lung and Circulation (2014), http://dx.doi.org/10.1016/j.hlc.2014.04.005

**Table 2** Types and frequencies of cardiac surgical procedures.

| SURGICAL SUBGROUP | n (%) |
|---|---|
| **ISOLATED CORONARY OPERATIONS** | **1038 (45.4)** |
| on-pump | 1002 (43.8) |
| off-pump | 36 (1.6) |
| **OTHER THAN ISOLATED CORONARY OPERATIONS** | **1249 (54.6)** |
| **single non-coronary** | **491 (21.5)** |
| AVR | 258 (11.3) |
| MVR | 60 (2.6) |
| mitral valve repair | 70 (3.1) |
| ascending aorta replacement or repair | 32 (1.4) |
| ASD closure | 46 (2.0) |
| cardiac myxoma | 25 (1.1) |
| **two procedures** | **576 (25.2)** |
| AVR+CABG | 214 (9.6) |
| AVR+mitral valve repair | 27 (1.2) |
| MVR+ tricuspid valve repair | 35 (1.5) |
| MVR+CABG | 83 (3.8) |
| CABG+mitral valve repair | 97 (4.2) |
| mitral+tricuspid valve repair | 47 (2.1) |
| AVR+MVR | 23 (1.0) |
| AVR+ascending aorta replacement or repair | 31 (1.4) |
| CABG+ascending aorta replacement or repair | 19 (0.8) |
| **three procedures** | **182 (8.0)** |
| AVR+CABG+mitral valve repair | 97 (4.2) |
| AVR+ CABG+tricuspid valve repair | 27 (1.2) |
| AVR+CABG+ascending aorta | 33 (1.4) |
| AVR+MVR+CABG | 25 (1.1) |

predicted a significantly higher ($p<0.01$) mortality than we observed. This discrepancy continued to exist even in the highest risk deciles. The only exception was the second decile. The overall O:E ratio was 0.66. The smoothed curve in Fig. 2 runs below the theoretical ''perfect-match'' reference line in all risk categories except for the very high mortality probabilities.

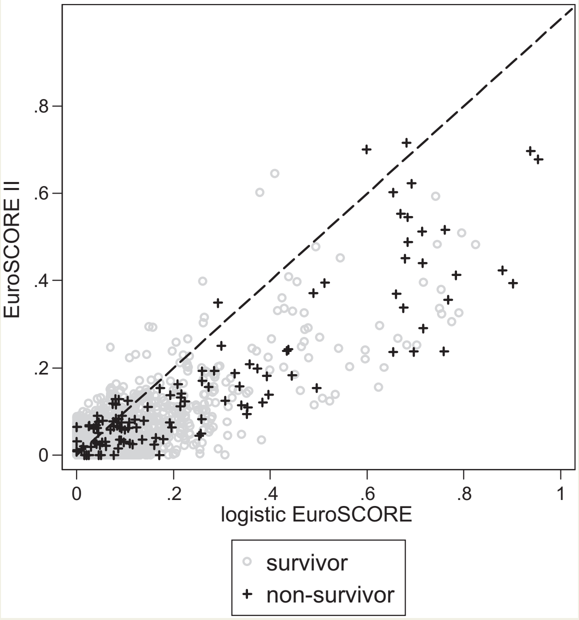

**Figure 1** Comparison of the expected mortality predicted by the logistic EuroSCORE (horizontal axis) and EuroSCORE II (vertical axis).

The Hosmer-Lemeshow test (Table 4) also revealed a significant deviation between expected and observed mortality ($p=0.0084$) when EuroSCORE II was used for the outcome prediction, but in this case the observed was slightly higher than the expected mortality (O:E ratio = 1.19). The smoothed curve in Fig. 3 gives a visual impression of the calibration of EuroSCORE II. It demonstrates that the difference between the observed and the predicted outcome is greatest among the highest risk individuals, where EuroSCORE II spectacularly underestimated the operative risk.

**Table 3** Logistic EuroSCORE, Hosmer-Lemeshow test results.

| Deciles | n | Observed mortality, n(%) | Expected mortality, n(%) | Probability interval | HL, chi2 |
|---|---|---|---|---|---|
| 1 | 249 | 2 (0.8) | 3.1 (1.2) | 0.009-0.015 | 0.37 |
| 2 | 242 | 6 (2.5) | 4.2 (1.8) | 0.015-0.021 | 0.74 |
| 3 | 214 | 0 (0.0) | 5.0 (2.3) | 0.021-0.025 | 5,14 |
| 4 | 211 | 2 (0.9) | 6.3 (3.0) | 0.026-0.033 | 3,01 |
| 5 | 229 | 6 (2.6) | 8.7 (3.8) | 0.033-0.043 | 0.87 |
| 6 | 233 | 6 (2.6) | 11.3 (4.9) | 0.043-0.055 | 2.62 |
| 7 | 223 | 12 (5.4) | 14.2 (6.4) | 0.055-0.072 | 0.35 |
| 8 | 229 | 18 (7.9) | 19.9 (8.7) | 0.073-0.103 | 0.20 |
| 9 | 229 | 14 (6.1) | 30.8 (13.4) | 0.103-0.175 | 10.54 |
| 10 | 228 | 57 (25.0) | 83.2 (36.5) | 0.175-0.908 | 12.96 |
| Total | **2287** | **123 (5.4)** | **186.6 (8.2)** | **0.009-0.908** | **36.81**<br>p=0.0001 |

Please cite this article in press as: Koszta G, et al. Performance of EuroSCORE II in Hungary: A Single-centre Validation Study. Heart, Lung and Circulation (2014), http://dx.doi.org/10.1016/j.hlc.2014.04.005

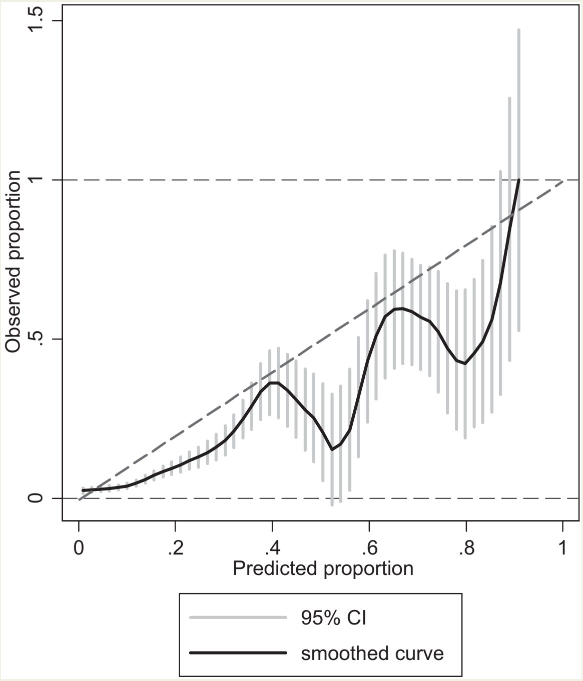

Figure 2 Calibration plot of the logistic EuroSCORE. The smoothed curve is the result of local polynomial regression using Epanechnikov kernel function with a bandwidth of 0.05.

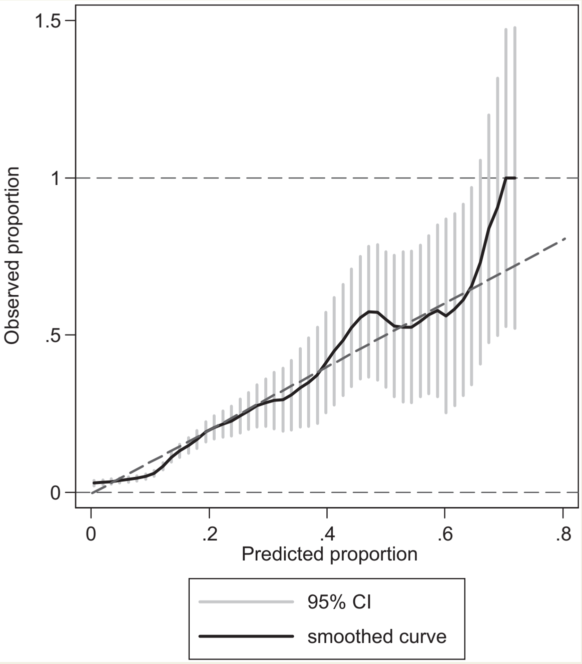

Figure 3 Calibration plot of EuroSCORE II. The smoothed curve is the result of local polynomial regression using Epanechnikov kernel function with a bandwidth of 0.05.

## Accuracy of the Logistic EuroSCORE and EuroSCORE II

The sensitivity and specificity of each potential cut-off point were calculated in ROC analysis for both scores (Fig. 4). The results showed an AUC of 0.8177 (95% CI: 0.7786-0.8569) for EuroSCORE II, which is slightly higher than the AUC (0.8017, 95% CI: 0.7596-0.8438) of the logistic EuroSCORE, the difference is not significant (p=0.1930).

The Brier score for EuroSCORE II was 0.0447, which is also non-significantly lower than the logistic EuroSCORE's Brier score of 0.0457 (p=0.775).

The Brier score was equal or higher than 0.25 in the case of 119 patients (5.2%) if EuroSCORE II was used for the outcome prediction. It was 126 (5.5%) with the logistic EuroSCORE. This means that the frequency of inaccurate predictions given by the old logistic EuroSCORE was only

Table 4 EuroSCORE II, results of the Hosmer-Lemeshow test.

| Deciles | n | Observed mortality, n(%) | Expected mortality, n(%) | Probability interval | HL, chi2 |
|---|---|---|---|---|---|
| 1 | 234 | 0 (0.0) | 1.6 (0.7) | 0.005-0.008 | 1.61 |
| 2 | 224 | 4 (1.8) | 2.1 (0.9) | 0.008-0.011 | 1.67 |
| 3 | 231 | 2 (0.9) | 2.8 (1.2) | 0.011-0.014 | 0.25 |
| 4 | 228 | 6 (2.6) | 3.6 (1.6) | 0.014-0.017 | 1.68 |
| 5 | 227 | 2 (0.9) | 4.6 (2.0) | 0.018-0.023 | 1.46 |
| 6 | 230 | 2 (0.9) | 5.9 (2.6) | 0.023-0.028 | 2.63 |
| 7 | 232 | 10 (4.3) | 7.5 (3.2) | 0.028-0.037 | 0.87 |
| 8 | 225 | 20 (8.9) | 9.9 (4.4) | 0.037-0.054 | 10.70 |
| 9 | 228 | 18 (7.9) | 16.3 (7.1) | 0.054-0.092 | 0.20 |
| 10 | 228 | 59 (25.9) | 48.9 (21.5) | 0.093-0.718 | 2.63 |
| Total | 2287 | 123 (5.4) | 103.2 (4.5) | 0.005-0.718 | 23.70<br>p=0.0084 |

Please cite this article in press as: Koszta G, et al. Performance of EuroSCORE II in Hungary: A Single-centre Validation Study. Heart, Lung and Circulation (2014), http://dx.doi.org/10.1016/j.hlc.2014.04.005

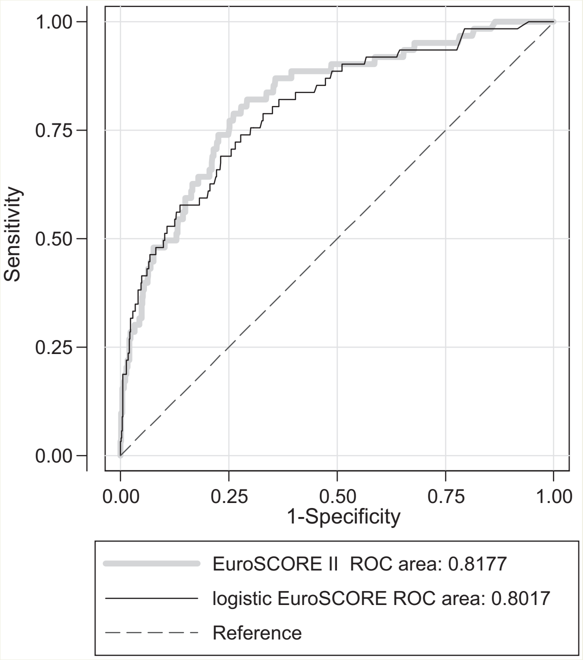

**Figure 4** Comparison of the logistic EuroSCORE's and EuroSCORE II's ROC curves.

non-significantly higher than that experienced with the new EuroSCORE (p=0.6521).

## Analysis of Cardiac Surgical Subgroups

Further analyses were made on the surgical subgroups created according to the ''weight of intervention'' categories defined by EuroSCORE II (Table 5). The result of the Hosmer-Lemeshow test revealed a good calibration for EuroSCORE II in the single coronary operation group (p=0.5789). The new EuroSCORE slightly over-predicted the risk (O:E ratio=0.75). Its accuracy was also excellent with an AUC of 0.8118 (95% CI: 0.7132-0.9105).

The calibration in the *''other than isolated coronary''* group showed less perfect goodness-of-fit with an HL-test p-value of 0.0084. The ROC AUC was also smaller: 0.7873 (95% CI: 0.7401 – 0.8345). This group unites extremely diverse interventions and patients (see Table 2) that can be divided into three further subgroups according to the complexity of the intervention.

In the ''single non-coronary'' group the HL-test indicated poor calibration. Observed mortality was 1.77 times higher than expected mortality. In the two combined operation groups the accuracy of EuroSCORE II diminished with the increasing complexity of the procedure and the

**Table 5** **Results of the subgroup analyses.**

| Surgical subgroup | N | Risk interval, % | Exp,n (%) | Obs, n (%) | O:E | HL, chi2 | HL, p | ROC AUC | 95% CI | BS, mean | BS ≥0.25, n (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ISOLATED CORONARY | 1038 | 0.5-46.9 | 29.4 (2.8) | 22 (2.1) | 0.75 | 8.52 | 0.5789 | 0.8118 | 0.7132-0.9105 | 0.0195 | 22 (2.1) |
| OTHER THAN ISOLATED CORONARY | 1249 | 0.5-71.8 | 73.7 (5.9) | 101 (8.1) | 1.37 | 23.70 | <0.001 | 0.7873 | 0.7401-0.8345 | 0.0649 | 97 (7.8) |
| single non-coronary | 491 | 0.5-65.5 | 17 (3.5) | 30 (6.1) | 1.77 | 31.060 | *0.0006 | 0.8750 | 0.7904-0.9595 | 0.0433 | 28 (5.7) |
| two procedures | 576 | 0.9-71.8 | 38.2 (6.6) | 45 (7.8) | 1.18 | 34.80 | *0.0001 | 0.7199 | 0.6358-0.8040 | 0.0649 | 45 (7.8) |
| three procedures | 182 | 2.2-53.0 | 18.6 (10.2) | 26 (14.3) | 1,4 | 34.80 | *0.0001 | 0.6144 | 0.4926-0.7362 | 0.1225 | 24 (13.2) |
| ELECTIVE PROCEDURES | 1649 | 0.5-41.7 | 49.3 (2.9) | 54 (3.2) | 1.1 | 13.54 | 0.1396 | 0.7679 | 0.6998-0.8360 | 0.0291 | 54 (3.3) |
| NON-ELECTIVE PROCEDURES | 593 | 0.8-71.8 | 53.9 (9.1) | 69 (11.6) | 1.28 | 14.18 | *0.0145 | 0.7913 | 0.7379-0.8448 | 0.0876 | 65 (10.9) |
| urgent procedures | 488 | 0.8-56.2 | 30.6 (6.3) | 32 (6.6) | 1.05 | 6.44 | *0.0921 | 0.7067 | 0.6246-0.7887 | 0.0606 | 32 (6.6) |
| emergency procedures | 77 | 1.9-65.5 | 14.6 (19.0) | 25 (32.5) | 1.71 | 12.63 | **0.0055 | 0.7708 | 0.6602-0.8814 | 0.1945 | 21 (27.3) |
| salvage procedures | 28 | 6.2-71.8 | 8.8 (31.3) | 12 (42.9 | 1.36 | 9.39 | **0.0245 | 0.5626 | 0.3269-0.7986 | 0.2649 | 12 (42.86) |
| TOTAL | 2287 | 0.5-71.8 | 103.2 (4.5) | 123 (5.4) | 1.19 | 23.70 | 0.0084 | 0.8177 | 0.7786-0.8569 | 0.0443 | 119 (5.2) |

Exp: expected mortality.
Obs: observed mortality.
O/E: observed-expected mortality ratio.
HL: Hosmer-Lemeshow test.
ROC AUC: receiver operating characteristics, area under the curve.
CI: confidence interval.
BS: Brier score.
*three groups in HL-test
**five groups in HL-test

**Table 6** Cross references among the subgroups.

| Weight of intervention | Urgency | | | | |
|---|---|---|---|---|---|
| | elective | urgent | emergency | salvage | total |
| CABG | | | | | |
| n | 723 | 277 | 36 | 2 | 1038 |
| O | 4 | 8 | 10 | 2 | |
| E | 13.1 | 9.9 | 5.9 | 0.5 | |
| single non-CABG | | | | | |
| n | 406 | 53 | 18 | 14 | 491 |
| O | 10 | 4 | 8 | 8 | |
| E | 8.4 | 2.4 | 3.2 | 3.1 | |
| two procedures | | | | | |
| n | 445 | 102 | 17 | 12 | 576 |
| O | 24 | 10 | 7 | 4 | |
| E | 18.5 | 10.3 | 4.2 | 5.2 | |
| three procedures | | | | | |
| n | 120 | 56 | 6 | 0 | 182 |
| O | 16 | 10 | 0 | 0 | |
| E | 9.4 | 7.9 | 1.32 | 0 | |
| total | 1694 | 488 | 77 | 28 | 2287 |

calibrations were also unsatisfactory. However, the deviations were smaller compared with the single non-coronary procedures.

The Brier score was lowest in the isolated CABG group 0.0195 with the lowest frequency (2.1%) of the patients with higher than 0.25. The BS increased with the complexity of the intervention and reached a rather high level of 0.1225 among those patients who underwent the most complex operations, indicating that the risk prediction was ambiguous in 13.2% of the patients in this group.

## Analysis of the Procedural Urgency Subgroups

The O:E ratio of the elective procedures was 1.1 with no significant deviation between observed and expected mortality (HL-test p=0.1396). Both the ROC analysis and the BS revealed good discriminative power and accuracy in this subgroup.

In the urgent operation group the O:E ratio of 1.045 showed only a small difference and the result of the HL-test was not significant (p=0.0921). The ROC AUC reached only the acceptable level; however, the Brier scores were equal or higher than 0.25 in only 6.6% of the patients, which is slightly higher than this ratio in the whole cohort.

In the emergency and the salvage groups neither the calibration nor the accuracy proved to be satisfactory with very high ratios of ambiguous predictions. We can observe a discrepancy between the result of ROC analysis and the Brier score in the emergency group, where the ROC AUC was as high as 0.7708, but the mean BS was 0.1945 with a ratio of 27.27% for BS≥0.25.

## Cross References Between the Subgroups

Table 6 contains cross references among the subgroups. It can be seen that, in general, the O:E ratios increase with the degree of urgency. This Table also reveals the causes of the surprisingly high O:E ratio in the single non-coronary group by revealing that the ratios of the emergency and the salvage operations were disproportionally high compared with the other subgroups.

# Discussion

The data collection for this validation study began after the completion of the EuroSCORE II project's recruitment phase. Consequently, these results are probably free from the effect of the continuous performance decline described previously by Hickey et al [6].

We have not found major differences in the composition of the risk predictors compared with the reported data in the original EuroSCORE II paper [1]. The mean age of our cohort was less than two years older and the ratio of female participants was slightly higher. No clinically meaningful difference was found in the ratio of elective and non-elective operations.

In accordance with other validation studies we have proved that EuroSCORE II is more precise in the outcome prediction than its predecessor [2]. The new EuroSCORE effectively and reliably predicted the risk for the majority of the patients who underwent the most frequent coronary bypass procedures and proved to be satisfactory in general when the operation was performed electively. The O:E ratio of 0.75 indicated that EuroSCORE II tends to overestimate the real risk in coronary patients. However, it did not reach the level of statistical significance.

EuroSCORE II under-estimated the real risk in those patients with higher than 0.6 EuroSCORE II p-values. The broad and widening confidence interval on the smoothed calibration curve in the higher risk categories is another sign indicating the uncertainty of the prediction.

The second area of poor goodness-of-fit was that of the emergency and the salvage procedures. The unsatisfactory calibration in the single non-coronary group is probably partially the consequence of the disproportionally higher ratio of emergency and salvage operations in this group. The other possible explanation is the older age of the patients who underwent aortic valve replacement. They were approximately five years older than the mean of the cohort.

The incidence of patients with a BS equal or higher, than 0.25 was also calculated. We interpreted this ratio as a marker of uncertainty of prediction. It was fairly low: 2.1% among the coronary patients and 3.3% in the elective procedures. However, our results showed that more than 10% of the patients received uninformative or even misleading forecasts regarding the probability of death within the postoperative 30 days when non-elective procedures were needed. This ratio was more than one-quarter among the patients who underwent emergency operations and only a little lower than half in salvage operations. The ROC area reached lowest level in the salvage operations. However, we must add here that the low number of patients in this cohort prevents us from drawing profound conclusions. These findings are in keeping with the results of the recently published paper by Grant et al [7] on the performance of the new EuroSCORE among emergency patients. Besides the inadequate accuracy, they have found an almost perfect O:E ratio in this group. However, EuroSCORE II overestimated the probability of mortality among the high-risk patients conversely compared with our own results.

The question may thus be raised as to whether these findings in the non-elective subgroups are due to the poor calibration of EuroSCORE II or whether other local and healthcare-related factors also have to be taken into account? When searching for an answer to the first part of the question, some practical aspects needed to be mentioned. EuroSCORE II gave inaccurate predictions when emergency operations were needed. This is understandable in situations where incomplete information collection may contribute to the more difficult prediction of the expectable outcome. It is also likely that the acutely deteriorated physiology results in more complex and multiple ways of the interactions among the risk factors. Unfortunately, interactions were not defined in EuroSCORE II. In addition to older age, those determinants that may modify the effect of other risk predictors most extensively probably account for the different degrees of urgency. Careful evaluation of the results provided by other studies may help exposing the local factors that otherwise cannot be distinguished from the inadequate calibration.

Table 7 summarises the results of some recently published validation studies [8–15]. It is clearly noticeable that the calibration of EuroSCORE II is not perfect when all types of procedures are evaluated together. The results of the subgroup analyses are also discordant.

In a large validation study [11] Chalmers et al found an acceptable goodness-of- fit in the isolated CABG group with an immediately not significant H-L p-value. In contrast to our results, the new EuroSCORE showed excellent performance in the isolated MVR group with a highly non-significant Hosmer-Lemeshow test and an almost perfect ROC AUC. It was even better than in the isolated AVR group.

A multi-centre validation study published by Bareli et al [8] concluded that the new version of EuroSCORE is not satisfactorily calibrated. Although it did well in the low risk categories, it increasingly over-predicted the risk of death among patients with higher than 0.4 EuroSCORE II p-values, a result which is contrary to our own findings. Concerns were also raised about the inclusion of non-significant risk factors that failed to increase the performance. The lack of a clinically important high risk condition such as post-infarct septal rupture was questioned.

Zhang et al [15] performed a validation study among patients who underwent valve surgery. EuroSCORE II's goodness-of-fit was excellent in the single valve surgery group, while it significantly underestimated the risk in the multiple valve group.

Another paper was published from China by Wang et al [14]. More than 11,000 other-than-coronary procedures were evaluated in multiple-centre setting. The EuroSCORE II non-significantly over-predicted the risk even in the combined operation groups.

Probably the largest validation study to date was that published by Grant et al [13]. The sample size was comparable to the original derivation dataset. Prospective data collection involved all cardiac surgical centres in the UK. This study found good accuracy, but poor overall goodness- of-fit. The subgroup analysis indicated inadequate calibration in the CABG group, where the new EuroSCORE II, similar to our results, overestimated the real likelihood of mortality. The calibration and the accuracy were excellent or satisfactory in the other surgical subgroups. Carnero-Alcázar et al [10] found very similar EuroSCORE II performance compared to the present study.

In a dataset from Liverpool [16], UK an O:E ratio of 1.38 was published. Similarly to our result their analysis clearly revealed the EuroSCORE II's tendency to under-predict the risk especially in the higher risk categories. The new EuroSCORE did not improve the risk prediction compared with its old version as reported in collaborative study from two European centres [17].

We are well aware of the limitations of our study, of which probably the greatest is the single-centre design. The sample size of our dataset is not small in general; however, it did not provide enough cases to perform an analysis of the specific operation types such as procedures on different heart valves and the ascending aorta. The relatively low number of

Table 7 Review of the main results of EuroSCORE II validation studies.

| author | year | country, setting | cohort | O:E ratio | HL, p | ROC AUC |
|---|---|---|---|---|---|---|
| Chalmers [10] | 2006-2010 | UK, single-centre | all procedures, n=5576 | 1.1 | <0.001 | 0.79 |
| | | | CABG, n=2913 | 1.12 | 0.052 | 0.79 |
| | | | AVR, n=814 | 1.1 | 0.07 | 0.69 |
| | | | MVR, n=340 | 0.71 | 0.6 | 0.87 |
| | | | AVR+CABG, n=517 | 1.0 | 0.38 | 0.74 |
| | | | aortic, n=351 | 1.21 | 0.43 | 0.81 |
| | | | misc.,n=642 | 1.84 | 0.99 | 0.7 |
| Barili [7] | 2006-2011 | Italy, multicentre | all procedures, n=12325 | 0.79 | <0.05 | 0.82 |
| Biancari [8] | 2006-2011 | Finland, single-centre | CABG, n=1027 | 1.2 | not reported | 0.852 |
| Di Dedda [11] | 2010-2011 | Italy, single-centre | all procedures, n=1090 | 1.21 | 0.22 | 0.81 |
| Carnero-Alcázar [9] | 2005-2010 | Spain, single-centre | all procedures, n=3798 | 1.27 | <0.001 | 0.851 |
| | | | CABG, n=1231 | 0.94 | 0.001 | 0.9 |
| | | | valvular, n=1727 | 1.39 | <0.001 | 0.827 |
| | | | combined, n=301 | 1.37 | 0.334 | 0.769 |
| | | | aortic, n=416 | 1.24 | 0.058 | 0.85 |
| | | | other, n=123 | 1.82 | 0.334 | 0.876 |
| Zhang [14] | 2006-2011 | China, single-centre | heart valve surgery, n=3479 | 1.28 | <0.0001 | 0.685 |
| | | | single valve, n=1106 | 1.03 | 0.103 | 0.792 |
| | | | multiple valve, n=2373 | 1.23 | <0.0001 | 0.605 |
| Grant [12] | 2010-2011 | UK, Ireland, multicentre | all procedures, n=23740 | 0.92 | 0.003 | 0.808 |
| | | | isolated CABG, n=12470 | 0.71 | 0.001 | 0.796 |
| | | | single non-CABG, 4984 | 1.06 | 0.398 | 0.781 |
| | | | two procedures, n=4766 | 1.04 | 0.108 | 0.731 |
| | | | three procedures, n=1520 | 0.96 | 0.044 | 0.733 |
| | | | isolated AVR, n=3116 | 0.81 | 0.319 | 0.772 |
| | | | AVR+CABG, n=2401 | 0.92 | 0.424 | 0.770 |

emergency patients in our cohort made it necessary to examine the results of many other centres.

## Conclusions

The new EuroSCORE indicated variably lower-than-real risk in the case of valve and combined surgeries and among high risk patients.

A perfectly calibrated preoperative risk stratification model is probably a theoretical illusion, because it also lacks the information of several not-included, ‘‘minor’’ factors [18] as well as the effect of intraoperative and postoperative events, which may additively modify the outcome. These individual or healthcare related factors may be systematic or sporadic. The former – if they have a negative impact on the patients' outcome – should be revealed and corrected during continuous quality control.

The effect of the intraoperative and early postoperative factors may be reflected most practically in the course of the postoperative parameters [19] and biomarkers such as cardiac troponines.

The new EuroSCORE can be considered a timely and necessary update of the former model with room left to improve on its performance. More frequent recalibration is essential to keep track of the changing practice [6]. Defining interactions between risk predictors in the logistic model may result in better performance [20], primarily in emergency procedures.

We think that the results of different cardiac surgical centres are comparable reliably on the basis of a universal EuroSCRE II risk model. However, an institutionally or regionally re-calibrated version of the original model may be necessary when used to facilitate the clinical decision making in those centres, where the characteristics of health care or the patients are different from the Western-European ones.

## Conflict of interest

None to declare

The authors of the study have not received any financial support

## Acknowledgements

The authors would like to say Thank You for Gergely Nagy MD, PhD for his critical remarks.

Please cite this article in press as: Koszta G, et al. Performance of EuroSCORE II in Hungary: A Single-centre Validation Study. Heart, Lung and Circulation (2014), http://dx.doi.org/10.1016/j.hlc.2014.04.005

## References

[1] Nashef SA, Roques F, Sharples L, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. Eur J Cardiothorac Surg 2012;41:734–44.

[2] 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.

[3] Farsang C, Naditch-Brule L, Perlini S, Zidek W, Kjeldsen SE. Inter-regional comparisons of the prevalence of cardiometabolic risk factors in patients with hypertension in Europe: the GOOD survey. J Hum Hypertens 2009;23:316–24.

[4] Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38.

[5] Gutierrez R, Linhart J, Pitblado J. From the help desk: Local polynomial regression and Stata plugins. Stata Journal 2003;3:412–9.

[6] Hickey G, Grant S, Murphy G, Bhabra M, Pagano D, McAllister K, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. Eur J Cardiothorac Surg 2013;43:1146–52.

[7] Grant S, Hickey G, Dimarakis I, Cooper G, Jenkins D, Uppal R, et al. Performance of the EuroSCORE models in emergency cardiac surgery. Circ Cardiovasc Qual Outcomes 2013;6:178–85.

[8] Barili F, Pacini D, Capo A, Rasovic O, Grossi C, Alamanni F, et al. Does EuroSCORE II perform better than its original versions? A multicentre validation study. Eur Heart J 2013;34:22–9.

[9] Biancari F, Vasques F, Mikkola R, Martin M, Lahtinen J, Heikkinen J. Validation of EuroSCORE II in patients undergoing coronary artery bypass surgery. Ann Thorac Surg 2012;93:1930–5.

[10] Carnero-Alcazar M, Silva Guisasola J, Reguillo Lacruz F, Maroto Castellanos L, Cobiella Carnicer J, Villagran Medinilla E, et al. Validation of EuroSCORE II on a single-centre 3800 patient cohort. Interact Cardiovasc Thorac Surg 2013;16:293–300.

[11] Chalmers J, Pullan M, Fabri B, McShane J, Shaw M, Mediratta N, et al. Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. Eur J Cardiothorac Surg 2013;43:688–94.

[12] Di Dedda U, Pelissero G, Agnelli B, De Vincentiis C, Castelvecchio S, Ranucci M. Accuracy, calibration and clinical performance of the new EuroSCORE II risk stratification system. Eur J Cardiothorac Surg 2013; 43:27–32.

[13] Grant S, Hickey G, Dimarakis I, Trivedi U, Bryan A, Treasure T, et al. How does EuroSCORE II perform in UK cardiac surgery; an analysis of 23,740 patients from the Society for Cardiothoracic Surgery in Great Britain and Ireland National Database. Heart 2012;98:1568–72.

[14] Wang L, Han QQ, Qiao F, Wang C, Zhang X, Han L, et al. Performance of EuroSCORE II in patients who have undergone heart valve surgery: a multicentre study in a Chinese population. Eur J Cardiothorac Surg 2013.

[15] Zhang G, Wang C, Wang L, Lu F, Li B, Han L, et al. Validation of EuroSCORE II in Chinese Patients Undergoing Heart Valve Surgery. Heart Lung Circ 2013;22:606–11.

[16] Kirmani B, Mazhar K, Fabri B, Pullan D. Comparison of the EuroSCORE II and Society of Thoracic Surgeons 2008 risk tools. Eur J Cardiothorac Surg 2013.

[17] Howell N, Head S, Freemantle N, van der Meulen T, Senanayake E, Menon A, et al. The new EuroSCORE II does not improve prediction of mortality in high-risk patients undergoing cardiac surgery: a collaborative analysis of two European centres. Eur J Cardiothorac Surg 2013.

[18] Head S, Osnabrugge R, Howell NJ, Freemantle N, Bridgewater B, Pagano D, et al. A systematic review of risk prediction in adult cardiac surgery: considerations for future model development. Eur J Cardiothorac Surg 2013;43:e121–9.

[19] Tamayo E, Fierro I, Bustamante J, Heredia-Rodriguez M, Monjas P, Maroto L, et al. Development of the Post Cardiac Surgery (POCAS) prognostic score. Crit Care 2013;17:R209.

[20] Hosmer D, Lemeshow S. Applied Logistic Regression, 2nd ed., New York: Wiley-Blackwell; 2000.