Editor’s Choice - External Validation of Models Predicting Survival After Ruptured Abdominal Aortic Aneurysm Repair

S.C. van Beeka, J.J. Reimerinka, A.C. Vahl b, W. Wisselink c, R.J.G. Peters d, D.A. Legemate a, R. Balm a,*, on behalf of the Amsterdam Acute Aneurysm Trial Collaborators e

WHAT THIS PAPER ADDS
In patients with a ruptured abdominal aortic aneurysm (RAAA), prediction models could support the decision to select surgical or conservative treatment. In the present study the prediction model “updated GAS” most accurately predicted death, and accuracy was further improved after recalibration. In future clinical practice, the predictions can be used for case mix comparison between hospitals and tailoring the prognosis for patients and relatives. However, the updated GAS was insufficiently accurate to identify patients who would die despite intervention. Therefore, future studies should aim to improve the identification of these high risk patients to support the decision to withhold intervention.

Objective: Prediction of survival after intervention for ruptured abdominal aortic aneurysms (RAAA) may support case mix comparison and tailor the prognosis for patients and relatives. The objective of this study was to assess the performance of four prediction models: the updated Glasgow Aneurysm Score (GAS), the Vancouver scoring system, the Edinburgh Ruptured Aneurysm Score (ERAS), and the Hardman index.

Design, materials, and methods: This was a retrospective cohort study in 449 patients in ten hospitals with a RAAA (intervention between 2004 and 2011). The primary endpoint was combined 30 day or in hospital death. The accuracy of the prediction models was assessed for discrimination (area under the curve [AUC]). An AUC >0.70 was considered sufficiently accurate. In studies with sufficiently accurate discrimination, correspondence between the predicted and observed outcomes (i.e. calibration) was recalculated.

Results: The AUC of the updated GAS was 0.71 (95% confidence interval [CI] 0.66—0.76), of the Vancouver score was 0.72 (95% CI 0.67—0.77), and of the ERAS was 0.58 (95% CI 0.52—0.65). After recalibration, predictions by the updated GAS slightly overestimated the death rate, with a predicted death rate 60% versus observed death rate 54% (95% CI 44—64%). After recalibration, predictions by the Vancouver score considerably overestimated the death rate, with a predicted death rate 82% versus observed death rate 62% (95% CI 52—71%). Performance of the Hardman index could not be assessed on discrimination and calibration, because in 57% of patients electrocardiograms were missing.

Conclusions: Concerning discrimination and calibration, the updated GAS most accurately predicted death after intervention for a RAAA. However, the updated GAS did not identify patients with a ≥95% predicted death rate, and therefore cannot reliably support the decision to withhold intervention.

© 2014 European Society for Vascular Surgery. Published by Elsevier Ltd. All rights reserved.
Article history: Received 5 March 2014, Accepted 11 October 2014, Available online 3 December 2014

Keywords: Aortic aneurysm, Abdominal, Aortic rupture, Decision support techniques, Prediction model, Ruptured abdominal aortic aneurysm

INTRODUCTION
The overall death rate in patients with a ruptured aneurysm of the abdominal aorta (RAAA) is approximately 74% (95% confidence interval [CI] 72—77%).1 In patients reaching the hospital and undergoing intervention, the death rate ranges between 24% and 49%.2—4 Surgeons have proposed distinguishing between those who would potentially benefit from surgery, and those in whom it might be better to withhold intervention, for example after cardiopulmonary resuscitation.5—7 In current clinical practice, the decision to
start surgical or conservative treatment is based on a fast evaluation of the patients’ clinical condition, the surgeon’s experience, and the wishes of the patient. It is a subjective interpretation of a harsh reality by the doctor, the patient, and the relatives. A prediction model is a more standardized and objective way to evaluate the chances of successful intervention and might be helpful at these moments of vital choices. Further benefits of prediction models lie in case mix comparison between hospitals and a tailored prognosis for patients and relatives.

Several models have been developed to predict death after intervention in patients with a RAAA: the Glasgow Aneurysm Score (GAS), the Vancouver scoring system, after intervention in patients with a RAAA: the Glasgow for patients and relatives.

MATERIALS AND METHODS

A retrospective study was conducted in all consecutive surgically treated patients with a RAAA in the Amsterdam ambulance region between May 2004 and February 2011. The present study was carried out as a sequel to the previously published Amsterdam Acute Aneurysm Trial. Other details and analyses of this cohort have been published previously. None of these previous studies aimed to validate prediction models for patients with a RAAA. The Amsterdam ambulance region covers an area of 1025 km² with 1.38 million inhabitants. During the inclusion period, care for patients with a RAAA was centralized in two university hospitals and one teaching hospital in cooperation with seven regional hospitals. All patients with a RAAA in all ten hospitals of the region were registered prospectively by the vascular surgeons, and included in the present study. Patients with a previous aortic reconstruction, a RAAA with associated trauma or aortoenteric fistula, were excluded. The primary endpoint was the combined 30 day or in hospital death rate. Compared with some previous validation studies of the prediction models, in hospital death was added to the definition; from a patients’ perspective the ultimate goal is survival and being discharged. Approval from a medical ethics committee was not needed because of the observational design. This study adhered to the STrengthening the Reporting of Observational studies in Epidemiology (STROBE) guidelines.

Updated GAS

The updated GAS score was calculated with the formula: age (years) + 7 for cardiac comorbidity (defined as previous history of myocardial infarction, cardiac surgery, angina pectoris or arrhythmia) + 10 for cerebrovascular comorbidity (defined as previous history of stroke or transient ischemic attack) + 17 for shock (defined as an in hospital systolic blood pressure <80 mmHg) + 14 for renal insufficiency (defined as a pre-operative serum creatinine >160 µmol/L) + 7 for OR (Fig. S1, online supplement).

Vancouver score

The Vancouver score was calculated with the formula: age (years)*0.062 + loss of consciousness (yes = 1/no = −1) *1.14 + cardiac arrest (yes = 1/no = −1)*0.6 (Fig. S2, online supplement).

ERAS

The ERAS score was calculated with the formula: +1 for best recorded in hospital Glasgow coma scale (GCS) <15, +1 for in hospital systolic blood pressure <90 mmHg, +1 for pre-operative hemoglobin level <5.6 mmol/L. A score of 0 or 1 corresponded with a predicted death rate of 30%, a score of 2 with a predicted death rate of 50%, and a score of 3 with a predicted death rate of 80%.

Hardman index

The Hardman index was calculated with the formula: +1 for age >76 years, +1 for in hospital loss of consciousness, +1 for a pre-operative serum creatinine >190 µmol/L, +1 for pre-operative serum hemoglobin level <5.6 mmol/L, +1 for electrocardiographic (ECG) ischemia (defined as ST segment depression greater than 1 millimeter or an associated T wave change determined by a senior cardiologist [RJGP]). A score of 3 or more corresponded with a predicted death rate of 100%.

Data collection and statistical analysis

Data were collected from the medical records by the first and second authors. Data entry was done using Microsoft Access 2003 (Microsoft Corporation, Redmond, WA, USA) using field limits, univariate and multivariate checks. A valid way of coping with missing values is by imputation. Missing data were imputed for the variables blood pressure, hemoglobin, creatinine, cardiac comorbidity, cerebrovascular comorbidity, resuscitation, loss of consciousness, and GCS. Multiple imputation was done creating ten datasets. Age, sex, renal and pulmonary comorbidity, death, and the above mentioned imputed variables were used as predictors in the imputation model. Baseline characteristics and prediction model scores are reported in both the original dataset and in the imputed datasets (Tables 1 and 2).

The statistical analysis and the imputation procedure were done using IBM SPSS Statistics 19.0 (SPSS Inc.,
Overall performance represents the squared difference between the predicted outcome and actual outcome, and was assessed using the Brier Score. The Brier Score should be as close to 0 as possible and the threshold for a non-informative model was calculated to be at 0.23. Discrimination is the ability of a model to distinguish between dying and surviving patients and was assessed using the area under the receiver operating characteristics curve (AUC). An AUC >0.70 was considered sufficiently accurate. Second, in the models with an AUC >0.70, the calibration of the predictions was determined. Calibration refers to the agreement between the predicted and observed death rate. Calibration was assessed by dividing all patients into five comparable quintiles: 0–20%, >20–40%, >40–60%, >60–80% and >80–100%. Because patients with equal predictions were categorized in the same quintile, the sizes of the quintiles differed slightly between the several prediction models. Subsequently, the mean predicted death rate per quintile was plotted with the corresponding mean observed death rate. In addition, the Hosmer-Lemeshow (HL) chi-square test was done to compare the observed and predicted death rates. In the HL test p < .05 reflects a significant difference between the predicted and observed death rate which is a poor calibration. Third, the models with an AUC >0.70 and an HL test p < .05 were recalibrated using the ‘calibration intercept method’. Fourth, a subgroup analysis was done in patients with a predicted death rate of ≥95% to assess the accuracy in high risk patients in whom withholding intervention might be considered. Fifth, the accuracy of the prediction models was determined in patients treated with EVAR. In this second subgroup analysis, the threshold of the Brier score for a non-informative model was calculated to be at 0.20.

RESULTS

Of 539 patients with a RAAA in the greater Amsterdam region, 66 did not have an intervention and 24 had to be excluded because of other reasons (Fig. 1). The reasons to refrain from intervention were predominantly shock or resuscitation with an expected low chance of survival (n = 20), patient or patient’s family decision (n = 17), or unknown (n = 17). The updated GAS, the Vancouver score, and the ERAS of these patients without intervention is shown in the online supplement (Table S3, online supplement). The baseline characteristics of the 449 patients included in the analysis are shown in Table 1. Sixty-nine patients were treated with EVAR and 380 patients were treated with OR. The death rate was 36% (160/449, 95% CI 31–40).

**Table 1. Baseline pre-operative characteristics.**

<table>
<thead>
<tr>
<th>Pre-operative variable</th>
<th>Original data Median (IQR)</th>
<th>Missing data % (Number)</th>
<th>Imputed data Median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years</td>
<td>76 (69–80)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Lowest in hospital SBP in mmHg</td>
<td>90 (70–125)</td>
<td>11 (48/449)</td>
<td>90 (70–125)</td>
</tr>
<tr>
<td>Hemoglobin at ER in mmol/L</td>
<td>7.0 (5.9–8.0)</td>
<td>1 (5/449)</td>
<td>7 (5.9–8.0)</td>
</tr>
<tr>
<td>Creatinine at ER in μmol/L</td>
<td>106 (86–133)</td>
<td>3 (14/449)</td>
<td>107 (87–134)</td>
</tr>
<tr>
<td>% (Number) Missing data % (Number)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male:female</td>
<td>80:20 (360:89)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Cardiac comorbidity</td>
<td>42 (184/435)</td>
<td>3 (14/449)</td>
<td>43 (191/449)</td>
</tr>
<tr>
<td>Cerebrovascular comorbidity</td>
<td>15 (67/433)</td>
<td>4 (16/449)</td>
<td>15 (69/449)</td>
</tr>
<tr>
<td>In hospital cardiopulmonary resuscitation</td>
<td>11 (46/429)</td>
<td>4 (20/449)</td>
<td>12 (52/449)</td>
</tr>
<tr>
<td>In hospital loss of consciousness</td>
<td>21 (81/388)</td>
<td>14 (61/449)</td>
<td>21 (96/449)</td>
</tr>
<tr>
<td>Best recorded Glasgow Coma Scale &lt;15</td>
<td>17 (63/372)</td>
<td>17 (77/449)</td>
<td>18 (82/449)</td>
</tr>
<tr>
<td>ECG ischemia</td>
<td>21 (40/192)</td>
<td>57 (257/449)</td>
<td></td>
</tr>
<tr>
<td>EVAR:OR</td>
<td>15:85 (69:380)</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

RAAA = ruptured abdominal aortic aneurysm; SD = standard deviation; SBP = systolic blood pressure; IQR = interquartile range; ER = emergency room; EVAR = endovascular aneurysm repair; OR = open repair, ECG = electrocardiogram.

**Table 2. The outcomes and distribution in the original dataset and the imputed dataset of the GAS, the updated GAS, the Vancouver score and the ERAS.**

<table>
<thead>
<tr>
<th>Prediction model</th>
<th>Original dataset</th>
<th>Missing data % (number)</th>
<th>Imputed dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Updated GAS, mean (±SD)</td>
<td>93 (±15)</td>
<td>14 (64/449)</td>
<td>93 (±15)</td>
</tr>
<tr>
<td>Vancouver, median (IQR)</td>
<td>3.10 (2.66–3.72)</td>
<td>14 (64/449)</td>
<td>3.10 (2.66–3.74)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% (number)</th>
<th>Missing data % (number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERAS score ≤1</td>
<td>79 (274/349)</td>
</tr>
<tr>
<td>2</td>
<td>17 (61/349)</td>
</tr>
<tr>
<td>3</td>
<td>4 (14/349)</td>
</tr>
</tbody>
</table>

SD = standard deviation; GAS = Glasgow Aneurysm Score; ERAS = Edinburgh Ruptured Aneurysm Score.
66%, the observed death rate was 55% (95% CI 44–65). After recalibration, the plot slightly improved, although there was still a statistically significant deviation between the predicted and observed risks (HL test $p = .04$, Fig. 3). In the quintile of patients with a mean predicted death rate of 60%, the observed death rate was 54% (95% CI 44–64) after recalibration.

Subgroup analysis to assess the accuracy in high risk patients showed that no patients had a predicted death rate $\geq 95%$. In patients treated with EVAR, the Brier Score was 0.17, the AUC was 0.78 (95% CI 0.66–0.90), and in the HL test $p$ was .18.

**Vancouver score**
The median Vancouver score was 3.10 (interquartile range 2.66–3.72) (Table 2). The Brier Score was 0.22 and the AUC was 0.72 (95% CI 0.67–0.77). With regard to calibration, in the quintile of patients with a mean predicted death rate of 33%, the observed death rate was 21% (95% CI 14–31%), and in the quintile of patients with a mean predicted death

$$\frac{1}{1 + \exp(-5.486 + 0.052 \times \text{updated GAS score})}$$

Figure 3. The formula to calculate the predicted death rate using the updated GAS after recalibration. GAS = Glasgow Aneurysm Score.
rate of 89%, the observed death rate was 62% (95% CI 52—71%). Hence, the calibration plot showed an overestimation of death (HL test \( p < .01 \)) (Fig. 3). After recalibration, this overestimation decreased minimally (HL test \( p < .01 \), Fig. 4). In high risk patients there was a significant overestimation of the observed risk by the recalibrated model. In the quintile of patients with a mean predicted death rate of 82%, the observed death rate was 62% (95% CI 52—71).

Subgroup analysis to assess the accuracy in high risk patients showed that of 21 patients with a predicted death rate ≥95%, 18 patients died. In patients treated with EVAR, the Brier Score was 0.19, the AUC was 0.77 (95% CI 0.64—0.90), and the HL test \( p = .03 \).

**ERAS**

The distribution of patients per ERAS outcome is shown in Table 2. The Brier Score was 0.23 and the AUC was 0.58 (95% CI 0.52—0.64). Calibration was not assessed because of an AUC <0.70. Subgroup analysis to assess the accuracy in high risk patients showed that no patients had a predicted death rate ≥95%. In patients treated with EVAR, the Brier Score was 0.20 and the AUC was 0.55 (95% CI 0.50—0.60).

**Hardman index**

In 57% (257/449), the pre-operative ECGs were missing. Therefore, the Hardman index was excluded from the analysis.

**DISCUSSION**

The present study shows that following intervention for a RAAA, the updated GAS predicted death most accurately for both discrimination and calibration. The present study expands on previous studies externally validating the updated GAS, the Vancouver score, and the ERAS. First, a cut-off value of patients was set in whom withholding intervention might be considered. In this way, it was aimed to assess the additional value of the prediction models in clinical practice. Second, the number of patients included (n = 449) was higher than the previous largest study (n = 201). Finally, the updated GAS and Vancouver score were recalibrated to improve accuracy in the era of EVAR. Because of the large number of missing ECGs, no definite conclusions could be drawn for the Hardman index.

**Decision making**

The decision to withhold intervention in patients with a RAAA can be very difficult. Only extremely reliable models can be useful in clinical decision making and in identifying patients in whom withholding intervention might be considered. For this purpose, a cut-off value for the predicted death rate was set at ≥95%. If the death rate was to be predicted accurately at 95%, the number needed to treat (NNT) would be 20. This cut-off value is arbitrary and could also have been 90% (NNT of 10) or 99% (NNT of 100). Different cut-off values can be used depending on the clinical situation. None of the prediction models met the criterion of identifying patients in whom to withhold intervention. This disappointing conclusion is in agreement with previous validation studies. Currently, the prediction models have insufficient accuracy to evaluate the chances of successful intervention and future studies should focus on improvement towards this aim. The usefulness of current prediction models lies in case mix comparisons between hospitals, and in a tailored prognosis for patients and relatives.

**Updated GAS**

The updated GAS predicted death most accurately for both discrimination and calibration. Several other studies have validated the GAS. In the only previous study including patients treated with EVAR, the AUC was 0.70 (95% CI 0.62—0.77). The calibration of the updated GAS was not assessed in this previous study. The strength of the previous validation (201 patients included, multicenter, prospective, including EVAR and OR) confirms the conclusion that the updated GAS is the most accurate in predicting death after intervention for a RAAA. If clinicians consider their patients to be comparable with those included in the present study, the model as shown in Fig. 3 can be used to predict the risk of dying after intervention.

**Vancouver score**

The Vancouver score discriminated sufficiently accurately, but even after recalibration its predictions still overestimated the death rate considerably. These results are in accordance with previous disappointing results on discrimination, but in conflict with previous fairly accurate results on calibration. Therefore, the accuracy of the Vancouver
score has not yet been proven and the present authors prefer the updated GAS.

**ERAS**

The prediction of death by the ERAS was insufficiently accurate. These results are in conflict with one validation study with sufficiently accurate discrimination, but in accordance with another validation study with insufficiently accurate discrimination. Concerning calibration, one previous validation reported an observed death rate of 50% in patients with a predicted death rate of 80% (estimated from figure). Because results regarding the ERAS are conflicting, the present authors question its precision.

**Hardman index**

The Hardman index was excluded from the analysis because data on one variable, presence of electrocardiographic (ECG) ischemia, were missing in 57% of patients. The missing ECGs are a drawback of the present study, and also of the scoring system. Most surgical trainees and vascular surgeons do not know how to interpret an ECG with sufficient precision to use it as a variable in a prediction model. From a cardiac perspective, acute ischemia defined by ST segment depression greater than 1 millimeter or an associated T wave change is an oversimplification of the great diagnostic value of an ECG. Based on these considerations, the present authors are convinced that in their clinical practice the contribution of a pre-operative ECG is limited and, consequently, that the Hardman index is not a useful prediction model.

**EVAR**

The predictions by the updated GAS and the Vancouver score appeared slightly more accurate in patients treated with EVAR compared with all patients. The accuracy of the ERAS appeared similar for patients treated with EVAR compared with all patients. Recently published randomized clinical trials reported comparable death rates after EVAR and OR. This indicates that the risk profiles are based on the same pre-operative variables and that the accuracy of the prediction models probably does not differ substantially between both interventions. Also, because of a low event rate (19/69) and wider confidence intervals in patients treated with EVAR, the present authors are reluctant to draw definite conclusions regarding the accuracy in patients treated with EVAR separately.

**Limitations**

An important limitation of the present study was the exclusion of the Hardman index. Two other prediction models have been described in the literature, the RAAA-physiological and operative severity score for enumeration of mortality and morbidity (RAAA-POSSUM) and the Vascular Study Group of New England (VSGNE) RAAA score. The RAAA-POSSUM was not included in the present study because of its complexity including chest X-ray examination, and hence low clinical applicability. The VSGNE RAAA score was not included in the present study because of the use of an intra-operative variable, thereby making predictions prior to the intervention impossible. In short, the authors consider the variables of the three excluded prediction models unsuitable for fast evaluation of the chances of successful intervention.

Another limitation of the present study was retrospective data collection. Probably, the variables ‘best recorded in hospital Glasgow coma scale’ for the ERAS and ‘loss of consciousness’ for the Vancouver score contain imprecise data. Another limitation was the amount of missing data (Tables 1 and 2). This is a consequence of the acute characteristic of the disease. The authors cope with this problem by multiple imputation. Death was included as a predictor in the imputation model to correct for the bias that the most missing data were in patients who died.

**Conclusions**

The updated GAS most accurately predicted death after intervention for a RAAA. However, the updated GAS did not identify patients with a predicted death rate ≥95%, and therefore cannot reliably support the decision to withhold intervention.

**ACKNOWLEDGEMENT**

The authors wish to thank Susan van Dieren for statistical support.

**CONFLICT OF INTEREST**

None.

**FUNDING**

Partial funding was provided by the Dutch Heart Foundation (project: 2002B197) and the AMC Foundation. The sponsor had no involvement in the study design, in the collection, analysis and interpretation of data; in the writing of the manuscript; and in the decision to submit the manuscript for publication.

**APPENDIX A. SUPPLEMENTARY MATERIAL**

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.ejvs.2014.10.012.

**REFERENCES**
