Eur J Vasc Endovasc Surg **34**, 514–521 (2007)
doi:10.1016/j.ejvs.2007.06.017, available online at http://www.sciencedirect.com on ScienceDirect

External Validity of a Mortality Prediction Model in Patients After Open Abdominal Aortic Aneurysm Repair Using Multi-level Methodology

V.G. Hadjianastassiou,[1*] P.P. Tekkis,[2] T. Athanasiou,[3] A. Muktadir,[4] J.D. Young[5] and L.J. Hands[6]

[1]*Specialist Registrar, Department of Vascular Surgery, 1st Floor, North Wing, St. Thomas' Hospital, Lambeth Palace Road, London SE1 7EH, UK,* [2]*Senior Lecturer & Consultant Surgeon, Department of Surgical Oncology & Technology, Imperial College of Science, Technology and Medicine, St. Mary's Hospital, London W2 1NY, UK,* [3]*Senior Lecturer & Consultant Cardiothoracic Surgeon, The National Heart and Lung Institute, Imperial College of Science, Technology and Medicine, St. Mary's Hospital, London W2 1PG, UK,* [4]*Senior Clinical Fellow, Oxford Transplant Unit, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK,* [5]*Consultant Anaesthetist & Director Adult Intensive Care Unit, John Radcliffe Hospital, Oxford, OX3 9DU, UK, and* [6]*Clinical Reader, Consultant Vascular Surgeon, Nuffield Department of Surgery, John Radcliffe Hospital, Oxford, OX3 9DU, UK*

Objectives. *Evaluation of the prognostic ability of the APACHE-AAA model in an independent group of post-operative (open) Abdominal Aortic Aneurysm (AAA) patients.*
Methods. *The model was applied to predict in-hospital mortality in 541 patients (325 elective and 216 emergencies; 489 from Oxford; 52 from Lewisham). Multi-level modelling was used to adjust for both the local structure and process of care and patient case-mix. Model performance was assessed using goodness-of-fit and subgroup analyses.*
Results. *The model's predictive ability to discriminate between dead and alive patients was very good (ROC area = 0.84). The model achieved a good fit across all strata of risk (Hosmer-Lemeshow C-test (8, N = 476) = 7.777, p = 0.456) and in all subgroups. The model was able to rank the ICUs according to their performance independently of the patient case-mix.*
Conclusion. *The APACHE-AAA model accurately predicted in-hospital mortality in a population of patients independent of the one used to develop it, confirming its validity. The multi-level methodology employed has shown that patient outcome is not only a function of the patient case-mix but instead predictive models should also adjust for the individual hospital-related factors (structure and process of care).*

Keywords: Hospital mortality; Intensive care units; Severity of illness index; Prognosis; Models, Statistical.

Introduction

The APACHE-AAA model,[1] based on the principles of the Acute Physiology and Chronic Health Evaluation (APACHE-II) methodology[2] was developed using data from 24 general Intensive Care Units (ICUs) in North Thames, United Kingdom (UK) collected over a 9 year period from 1992 to 2000. It was designed for use in applications of risk stratification modelling in the post-operative ICU care of abdominal aortic aneurysm (AAA) patients and so data were collected just after the completion of the operation, before ICU care had had any influence on outcome. It is the only published model which is applicable for both elective and emergency AAA repairs and was developed using a two-level multiple logistic regression analysis, by adjusting for any possible clustering of patients among the different ICUs (the so called ''ICU effect'' on patient outcome). It has been internally validated using measures of calibration, discrimination and subgroup analyses.

The model was created to fill a literature vacuum in risk stratification modelling in the post-operative AAA

*Corresponding author. Mr. V. G. Hadjianastassiou, DM(Oxon) FRCS (Gen) BMBCh(Oxon) BSc. Department of Transplantation, Guy's Hospital, Level 6, New Guy's House, St. Thomas' Street, London SE1 9RT, UK.
E-mail address: vassilis@doctors.org.uk

patient managed in the ICU. In its development study[1] it was advocated for use in comparative audit of the post-operative critical care facilities in this group of patients, by comparing the actual with the predicted mortality. It may also be used in ''outcome studies'' to determine the factors influencing AAA patient outcome in the post-operative ICU care setting. The model was also suggested to be useful as a guide for administrators when purchasing critical care services for AAA patients.[3] Furthermore, it may be used in research[4] involving AAA surgery to assess whether groups are comparable in terms of underlying case mix or may be used as a quantitative surrogate measure, summarising an AAA patient's post-operative clinical status, to aid exchange of information between clinicians. In a study[5] comparing the model with clinician predictions and those of neural networks, the APACHE-AAA was advocated for post-operative use as an adjunct in the process of generating informed prognosis to decrease uncertainty and promote communication among clinicians, patients and patients' families. Furthermore, the APACHE-AAA model has been shown[6] to be a more accurate risk stratification model than other contemporary models in its target population of post-operative AAA patients managed in ICU.

Despite the internal validity[1] of this model and its apparent success over other methods of prediction,[5,6] as the data used to develop the original model[1] were collected from a particular region of the UK, inferences about the applicability of the model elsewhere can not be made until it is validated in another region. The purpose of this study was to formally test the external validity of the APACHE-AAA model in a group of post-operative AAA patients managed in ICUs independent from the units used to develop the model, using multi-level methodology.

Methods

Data sources, study outcome and prognostic variables

Patient information was collected from post-operative AAA patients managed in the ICUs of John Radcliffe Hospital, Oxford (January 1999 to December 2004) and University Hospital Lewisham, London (October 2003 to September 2004), both in the UK. A comprehensive description of this external validation database has already been published,[6] with the most important feature being that there was no patient overlap with the original North-East Thames database used to develop the APACHE-AAA model.[1] These two databases comprised patients from different hospitals in separate geographical regions of the UK and as only the first admission to ICU was included for every patient, this ensured that the two databases were independent of each other. Eligible patients for inclusion in the study were those who underwent elective or emergency open surgical repair of AAA and were transferred post-operatively directly to the ICU for further management within the same hospital. Endovascular AAA repairs were excluded, in keeping with the original APACHE-AAA methodology.

The dataset definitions used in the development study of the APACHE-AAA[1] model were replicated in this study. The primary outcome was in-hospital mortality and the four prognostic variables used in the APACHE-AAA model include: (1) the Acute Physiology Score (APS) as defined in the original APACHE-II study[2] (2) the Chronic Health (CH) status, classified as a binary variable according to whether or not the patient had any chronic health dysfunction (that is, a history of severe organ system insufficiency or immunosuppression; (3) operative urgency (classified using the National Confidential Enquiry into Perioperative Deaths (NCEPOD) classification[7] of operations as an emergency (ruptured, leaking and symptomatic categorised by NCEPOD as emergency/urgent) or elective (NCEPOD scheduled/elective) surgical procedure; (4) chronological age. The values of these variables were taken to be the last recorded value on the anaesthesia sheet in the operating theatre after the end of the operation or the first recorded value on admission to the ICU (the latter for information on biochemical parameters), consistent with the APACHE-AAA methodology.[1]

Statistical methodology

Univariate logistic regression analysis was performed to confirm that the risk factors used in the APACHE-AAA model[1] were also associated with in-hospital mortality in this external validation group of patients. The variables whose univariate test had a $p < 0.25$ were entered into multiple regression analysis[8] to identify independent risk factors for in-hospital mortality, the influence of any interaction terms between the predictors, the influence of the year of operation and patient sex. Chi-square-for-trend analysis was performed to assess whether there was any variation in the ratio of emergency to elective cases over the years and to assess whether mortality changed during the study time period. Independent samples t-tests for continuous variables, and Pearson's chi-square tests for categorical variables, were performed to compare the original APACHE-AAA development (North-Thames) database with the external validation (Oxford/Lewisham)

database. Similar comparisons were made between the patients from Oxford and Lewisham hospitals.

The original APACHE-AAA model[1] was applied to the external validation database by using the published equations of the model (Table 1). Patients from the same ICU share many unmeasured prognostic factors (such as staffing levels and organisational features of the local ICU) resulting in the possibility that the outcome data of patients in the same ICU may correlate (that is, exhibiting ''clustering'' or sharing the same ''ICU effect''). In order to allow a true assessment of whether the model's predictor variables accurately described the patient case-mix, independently from the ICU effect, multilevel modelling[9,10] was applied. This method can mathematically adjust for the individuality of this structure and process of care of the external validation ICUs by using a different constant K for each ICU. This constant represented the individual ''ICU effect'' independent of the influence of case-mix on the patient outcome. Therefore, a multilevel model using all 26 units (24 from the development database and 2 from the validation database) was then applied to assess the ''ICU effect'' of each unit on patient outcome, and hence confirm the size of constant necessary to be used for each ICU unit in the equations (Table 1). The Gibb's re-sampling method[9] with 10 000 iterations was applied to reduce ''over-fitting'' (the generalisation error) and calculate confidence limits and correct bias in the parameter estimation.

The external validity of the APACHE-AAA model was evaluated by measures of calibration and discrimination, as well as subgroup analysis. Statistical analysis was two-sided using a significance level of $p < 0.05$. Calibration[8] or how well the model ''fits'' the observed outcome (goodness of fit) refers to the ability of the model to assign the correct probabilities of outcome to individual patients. This ability was assessed using the Hosmer-Lemeshow C statistic[11] in which a high p value would indicate a good model fit. Model discrimination refers to the ability of the model to be applied in any pair of patients, one of which goes on to die and the other lives on, and to assign higher probabilities of death to the patient who actually dies than the one who lives. This was evaluated by the C-Index which is equivalent to the area under the receiver operating characteristic (ROC) curve.[11]

Table 1. Equations used to apply the APACHE-AAA model in the Oxford/Lewisham population

(1) $\text{Logit}(p) = (0.05\ \text{Age}) + (0.13\ \text{APS}) + (1.58\ \text{Emergency}) + (0.36\ \text{CH dysfunction}) + \text{K}$

(2) $p = \exp^{\text{logit}(p)}/[1 + \exp^{\text{logit}(p)}]$

Logit(p) = the natural logarithm of the odds ratio of in-hospital death.
p = the probability of in-hospital death.
K = a constant with a range of possible values, defined by the 24 ICUs used to develop the APACHE-AAA model[1], with a mean (SD) value of −6.96 (0.70).
APS = Acute Physiology Score.
CH = Chronic Health.

Software

Analysis was performed using the computer software: ''Statistical Package for the Social Sciences'' version 12 for Windows® (SPSS, Chicago, Illinois, USA), and Intercooled STATA™ 8.0 for Windows (STATA corporation, College Station, TX, USA).

Results

Demographics and basic comparative data

An extensive description of the external validation database has already been published in a recent study[6] from our group. The database comprised 489 patients from Oxford and 52 patients from Lewisham. 65 patients from Oxford who had a missing CH status were included in the analysis. Analysis of the missing CH status data did not reveal a statistically significant bias of distribution of missing values among the categories of operative urgency and their associated mortality. In-hospital mortality rates were: for elective surgery patients ($N = 325$) 6.2% (95% confidence intervals (C.I.): 3.5−8.8%) and for emergency surgery patients ($N = 216$) 28.7% (95%C.I.: 22.5−34.9%). The mean (SD) age was 71.1 (8.0) years, 86.3% of all the patients were male and 12.2% had Chronic Health dysfunction. Patient sex ($p = 0.783$) and year of operation ($p = 0.308$) were not found to be independent predictors of outcome when adjusted for Age, APS, CH status and operative urgency. None of the six pairs of possible variable interaction terms were found to be independent predictors when adjusted for these 4 variables. There was a significant drop in the ratio of emergency to elective operations over the 6 years of study (Chi-Square for trend: (1, $N = 541$)=7.713, $p = 0.006$). There was no significant change in the mortality risk over this period (Chi-Square for trend: (1, $N = 541$)=0.794, $p = 0.373$).

A comparison of the external validation database to the original North Thames database, in terms of some of the important determinants of the patient case-mix and outcome is available in Table 2. The differences

shown for the APS and Age remain statistically significant in both elective and emergency patients. The corresponding results for the comparison between the Oxford and Lewisham patients are shown in Table 3.

Adjustment for the structure and process of care (multilevel modelling)

The 26 unit multilevel logistic regression model is shown in Table 4, demonstrating that the regression coefficients of the predictor variables did not change significantly in relation to the development model,[1] confirming the reliability of the model. The size of the adjustment to the constant necessary to compensate for the individual ''ICU effect'' for each of the 26 units, is shown in Fig. 1. This confirmed an ''outlier'' performance for the Oxford ICU (performing better than predicted) and hence the significant adjustment (-1.29 ± 0.36 (SD) from a mean (SD) ICU constant of -7.22 (0.51)) necessary to be applied to eliminate the better than ''average'' ICU performance before assessing whether the model adequately described the patient case-mix. Fig. 2 depicts the predicted in-hospital mortality (mean and 95% C.I.) for each of the 26 ICUs, with adjustment only for the patient case-mix (without adjustment for the individual ''ICU effect''). The outlier performance of Oxford is clearly shown with the C.I. of the predicted mortality (on the basis of patient case-mix alone) lying separately to the observed. In Fig. 3, after adjustment for both the patient case-mix and the individual ''ICU effect'' the model is shown to fit well all the ICUs, confirming that the model is valid for all ICUs, even in the case of an outlier performance. This multilevel logistic regression model, demonstrated a statistically significant variation in outcome between ICU units, as evidenced by the level two variance ($\sigma_u^2 = 0.223$, SD 0.100, Chi-Squared = 4.947, 1df, $p = 0.026$), due to the outlier performance of the Oxford ICU. This is in contrast to the development model[1] which did not show a significant difference between the ICU ''effects'' of the different hospitals ($\sigma_u^2 = 0.054$, SE 0.058, Chi-Squared = 0.833, 1df, $p = 0.361$).

Table 2. Comparison of the development and validation databases

Variable for comparison between databases		North-Thames (Development)	Oxford-Lewisham (Validation)	*P*-value
In-hospital mortality (%)	All patients	21.5	15.2	0.001[1]
	Elective	9.6	6.2	0.050[1]
	Emergency	46.9	28.7	<0.001[1]
APS −mean (SD) points	All patients	8.0 (6.7)	13.8 (6.8)	<0.001[2]
Age −mean (SD) years	All patients	71.1 (8.0)	73.0 (7.3)	<0.001[2]
Emergency workload (%)		31.9	39.9	0.001[1]
% of patients with CH dysfunction	All patients	24.0	12.2	<0.001[1]
	Elective	24.0	7.9	<0.001[1]
	Emergency	24.0	18.2	0.093[1]
Male patients (%)		83.5	86.3	0.119[1]
ICU stay − mean (SD) days		14.1 (4.5)	3.4 (6.8)	<0.001[3]

[1] Pearson's Chi-Square test.
[2] Independent samples t-test for unequal variances.
[3] Independent samples t-test for equal variances.

Discrimination, calibration and subgroup analyses

The discrimination properties of the APACHE-AAA model applied to the external validation database were well maintained with a C-Index (0.842, 95%C.I. = 0.799−0.885), which was not significantly different (Chi-Square test for two independent proportions (1, $N = 1751$, 476) = 0.022, $p = 0.881$) to that of the APACHE-AAA development study (0.845, 95% C.I. = 0.821−0.868). After adjustment for the individual ICU effect (and thereby eliminating the confounding effect of the structure and process of care on outcome), the APACHE-AAA equation was successful in achieving a good fit of the model to the external validation population. This confirmed that the model's predictor variables accurately described the patient case-mix. The model's calibration properties are shown in Fig. 4 depicting relatively uniform model predictions across all deciles of risk, as evidenced by the Hosmer-Lemeshow C-test (8, $N = 476$)=7.777, $p = 0.456$).

As a further measure of the internal validity of this model, subgroup analysis was performed, with respect to the chronic health status (Pearson's Chi-Square (1, $N = 476$)=6.266, $p = 0.617$) and the urgency of the operation (Pearson's Chi-Square (1, $N = 476$)=1.023, $p = 0.998$) respectively. The model's predictions rested within the 95% C.I. of the observed mortality across both categories of operative urgency and chronic health status.

Discussion

This study evaluated the application of the internally valid APACHE-AAA model, developed from a group of 24 ICUs in North Thames, to patients derived from two hospitals independent from the development group. The ability of the model to discriminate between survivors and patients who died was not significantly different compared to the original development

Table 3. Comparison of the Oxford and Lewisham constituent groups of the external validation database

Variable for comparison between hospitals		Oxford	Lewisham	*P*-value
In-hospital % mortality (N Dead/Total)	All patients	14.3 (70/489)	23.1 (12/52)	0.094[1]
	Elective	5.9 (17/288)	8.1 (3/37)	0.486[2]
	Emergency	26.4 (53/201)	60.0 (9/15)	0.014[2]
APS −mean (SD) points	All patients	13.8 (6.7)	13.7 (7.4)	0.908[3]
Age −mean (SD) years	All patients	73.1 (7.4)	72.5 (6.8)	0.611[3]
Emergency workload (%)		41.1 (201/489)	28.8 (15/52)	0.086[1]
% of patients with CH dysfunction	All patients	13.7 (58/424)	0.0 (0/52)	0.001[2]
	Elective	9.1 (22/241)	0.0 (0/37)	0.054[2]
	Emergency	19.7 (36/183)	0.0 (0/15)	0.078[2]
Male patients (%)		87.3 (427/489)	76.9 (40/52)	0.038[1]

[1] Pearson's Chi-Square test.
[2] Fisher's Exact Test.
[3] Independent samples t-test for equal variances.

study, with the ROC area[11] (C-index) remaining above 0.84. Adjustment for both patient case-mix and the local ''ICU effect'' allowed the model to fit the external validation population accurately across all risk strata and subgroups, confirming the reliability of the APACHE-AAA model even in an outlier ICU.

Data validity

The in-hospital mortality rate was consistent with the literature,[12,13] and the finding that gender was not a significant predictor of outcome was also consistent with previously published UK work.[14,15] The proportion of missing data in the study compared well with rates in similar studies, such as the original APACHE II study,[2] the APACHE-AAA study,[1] the UK APACHE II study[16] and the recent Vascular Biochemistry and Haematology Outcome Model for vascular surgery from the National Vascular Database.[17] Most importantly, analysis of the missing data of the CH status did not reveal a bias of distribution of missing values among the categories of operative urgency and their associated mortality.

On comparison of the two databases it was apparent that the significant difference in the in-hospital mortality between them was mostly due to the improved survival of the emergency cases in the external validation database. In addition, the proportion of emergency cases operated in this database was higher than in North Thames. The patients in the external validation database had a higher APS score at the end of their operation while the proportion of patients operated with CH dysfunction in the external validation group was half of the one in North-Thames. This latter difference was attributed mostly to the elective patients as there was no significant difference in the CH status of emergency patients between the two databases. Oxford and Lewisham patients were mostly comparable except for the significantly lower prevalence of CH dysfunction in Lewisham patients and the higher in-hospital mortality of their emergency patients.

Table 4. The 26 ICU APACHE-AAA Multilevel Model

Risk Factors	Coefficient β	SD	Odds ratio	95% C.I.
Age (per year)	0.05	0.01	1.05	1.04–1.07
Acute Physiology Score (per unit)	0.13	0.01	1.14	1.12–1.15
Emergency operation	1.58	0.07	4.83	4.21–5.55
Chronic Health dysfunction	0.40	0.07	1.49	1.29–1.71
Constant	−7.22	0.51		

Model interpretation

Many researchers believe that prognostic indices inevitably perform less well[18] when tested in an independent population. Specifically, the usual pattern in external validation studies is that the discriminatory property of the model is preserved at the expense of imperfect calibration.[16,19–21] A marked deterioration in the discriminative ability of a model cannot be corrected,[22] while poor calibration is easily correctable.[23]

There are various explanations in the literature to account for the deterioration of a model's performance in an external validation study. The factor with the biggest impact is the difference in the structure and process of care in the different institutions[22,24] which has an effect on patient outcome independently of the case-mix. The APACHE-AAA model had originally been developed by using multilevel methodology to enable it to have a range of possible equation constants, each one reflecting this individual ICU ''effect'' on the model's outcome. This ''ICU effect'' incorporates the combined influence of the hospital-related structure and process of care (such as organisation, financing, staffing levels, teamwork, volume and pressure of work, ICU

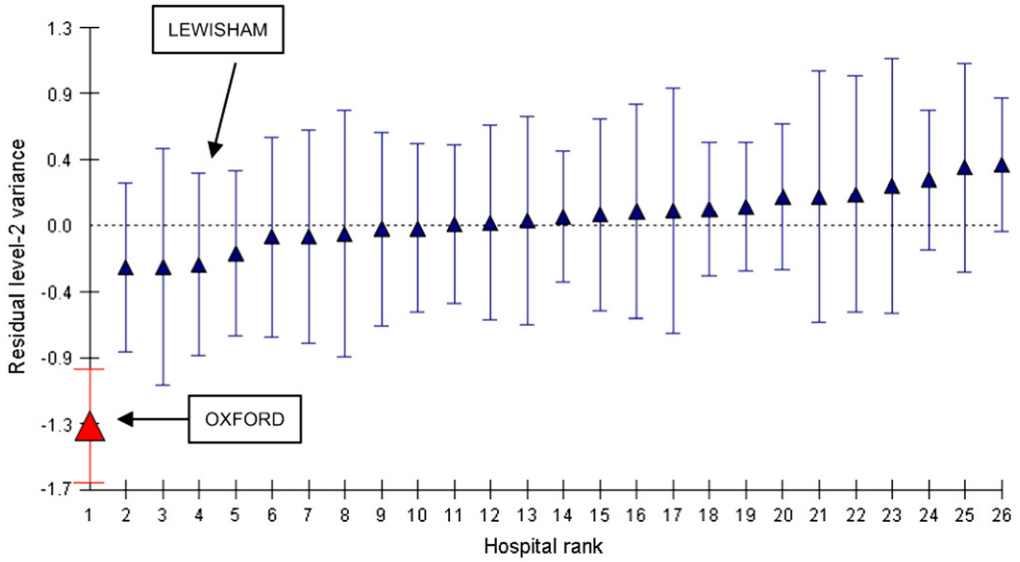
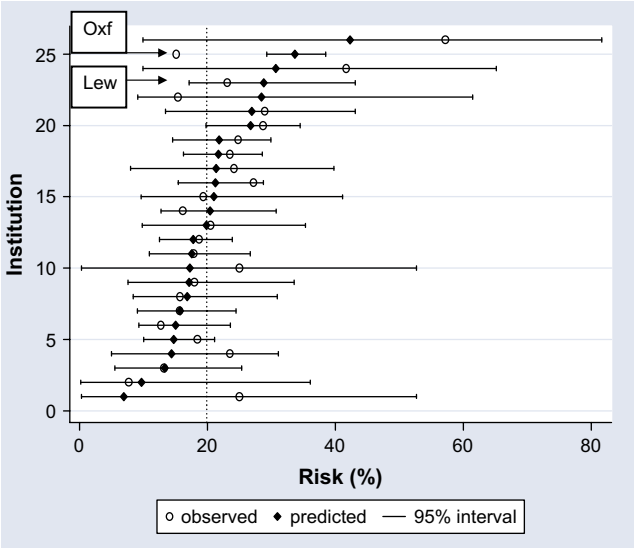


Fig. 1. Hospital ICU ranking according to the magnitude of the "ICU" effect on patient outcome.

admission and discharge policies, technology and implementation of pathways of care) on patient outcome, independently of the case-mix of the patient population, and can therefore be a confounding factor which has to be adjusted for. Contemporary models in vascular surgery only take into account the patient case-mix and ignore this "hospital effect". Another factor quoted in the literature is a temporal change in the relationship between the predictor variables and mortality with time due to medical progress.[25] This is unlikely in the present study as there was no significant change in the mortality rate across the 6 years of study. A discrepancy in the local referral patterns[18] and selection strategies (part of the individual



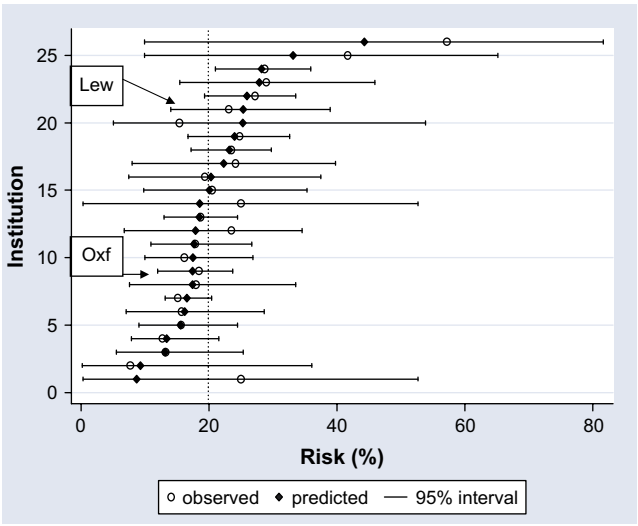
Legend:

Dotted line represents the mean population in-hospital mortality for all 26 ICUs

Oxf = Oxford ICU

Lew = Lewisham ICU

Fig. 2. Observed and predicted (95% C.I.) in-hospital mortality using the 26 unit multilevel logistic regression model, without adjusting for the "ICU effect".



Legend:

Dotted line represents the mean population in-hospital mortality for all 26 ICUs

Oxf = Oxford ICU

Lew = Lewisham ICU

Fig. 3. Observed and predicted (95% C.I.) in-hospital mortality using the 26 unit multilevel logistic regression model, after adjusting for the "ICU effect", showing good model prediction in both ICU of the external validation database.

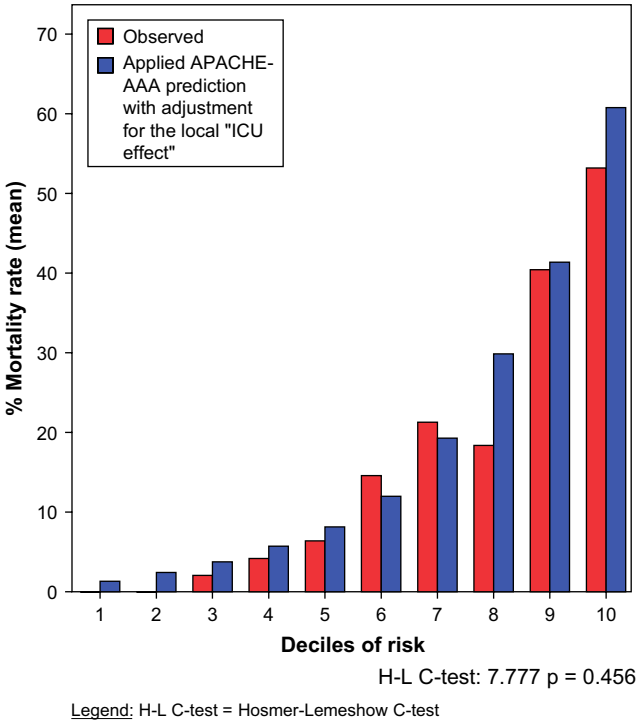


Fig. 4. Calibration bar chart of the applied APACHE-AAA model (after adjustment for the local ''ICU effect'') on the external validation database.

ICU process of care) can produce populations that are prognostically distinct, introducing a case-mix selection bias compared to the original database. A predictive model which accurately describes case-mix by its predictor variables should not be affected by any selection bias in a patient population, except for the case when the original model was ''over-fitted'' to its development population, by using too many variables resulting in a Type I error. This factor does not apply to this study as the rule of at least ten outcome events per potential predictor variable[23] was followed during development of the APACHE-AAA model to prevent this happening.

Multi-level methodology has overcome the main hurdle in applying a model to an independent patient population. The study presented here takes advantage of this method of modelling which allows adjustment for both case-mix and for the individual ICU effect (structure and process of care) on predicted outcome. This is an essential prerequisite for applying the principle of a ''one-model-fits-all'', as long as the individual hospital-related effects have been adjusted for.

Acknowledgements

We are grateful to: colleagues at Oxford and Lewisham hospitals who have contributed patients to this study; the Library staff at University Hospital Lewisham and especially Mrs Jane Coyte, for their help in providing articles from literature searches; Mr. S. Noel Clinical Nurse Specialist, Adult Intensive Care Unit, John Radcliffe Hospital, Oxford for his invaluable work with data collection.

The use of the term APACHE refers to the principles of the APACHE methodology as applied in Intensive Care patients and does not imply any collaboration with the developers of the APACHE software.

References

1 Hadjianastassiou VG, Tekkis PP, Hands LJ, Goldhill DR. Quantification of Mortality Risk after Abdominal Aortic Aneurysm repair. *Br J Surg* 2005;**92**:1092–1098.
2 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985;**13**: 818–829.
3 Hadjianastassiou VG, Tekkis PP, Poloniecki JD, Goldhill DR. Risk stratification in Abdominal Aortic Aneurysm surgery. The solution to HDU case selection? *Br J Surg* 2002;**89**:29.
4 Gunning K, Rowan K. Outcome data and scoring systems in ABC of intensive care. *BMJ* 1999;**319**:241–244.
5 Hadjianastassiou VG, Franco L, Jerez JM, Evangelou IE, Tekkis PP, Goldhill DR *et al*. Optimal prediction of mortality after abdominal aortic aneurysm repair with statistical models. *J Vasc Surg* 2006;**43**:467–474.
6 Hadjianastassiou VG, Tekkis PP, Athanasiou T, Muktadir A, Young JD, Hands LJ. Comparison of mortality prediction models after abdominal aortic aneurysm repair. *Eur J Vasc Endovasc Surg* 2007;**33**:536–543.
7 National Confidential Enquiry into Patient Outcome and Death. Classification of Operations (NCEPOD definitions). http://www.ncepod.org.uk/2004report/appendices.a.htm (accessed 12th May 2007).
8 Hosmer JA, Lemeshow S, eds. *Applied logistic regression*. 2nd ed. New York: Wiley & Sons, Inc; 2000.
9 Rasbash J, Browne W, Goldstein H, Yang M, Plewis I, Healey M *et al*, eds. *A User's Guide to MLwiN*. London: University of London; 2001.
10 Spiegelhalter DJ, Aylin P, Best NG, Evans SJW, Murray GD. Commissioned analysis of surgical performance by using routine data: lessons from Bristol inquiry. *J R Stat Soc A* 2002;**165**:1–31.
11 Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–387.
12 Ashley S, Riddler B on behalf of the Audit and Research Committee of the Vascular Surgical Society of Great Britain and Ireland. National Vascular Database Report 2002. Dendrite Clinical Systems Ltd, April 2003 http://www.vascularsociety.org.uk/Docs/NVD2002.pdf. (accessed 12th May 2007).
13 National Confidential Enquiry into Patient Outcome and Death. Abdominal Aortic Aneurysm: a Service in need of surgery? http://www.ncepod.org.uk/2005report2/Downloads/summary.pdf (accessed 12th May 2007).
14 Evans SM, Adam DJ, Bradbury AW. The influence of gender on outcome after ruptured abdominal aortic aneurysm. *J Vasc Surg* 2000;**32**:258–262.
15 Earnshaw J, Riddler B on behalf of the Audit and Research Committee of the Vascular Surgical Society of Great Britain and Ireland. National Vascular Database Report 2001. Dendrite Clinical Systems Ltd, May 2001 http://www.vascularsociety.org.uk/Docs/Dendrite_VSSGBI.pdf (accessed 12th May 2007).
16 Rowan KM, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland-II: outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method. *BMJ* 1993;**307**:977–981.

17 Prytherch DR, Ridler BM, Ashley S. Risk-adjusted predictive models of mortality after index arterial operations using a minimal data set. *Br J Surg* 2005;**92**:714–718.
18 Charlson ME, Ales KL, Simon R, MacKenzie CR. Why predictive indexes perform less well in validation studies. Is it magic or methods? *Arch Intern Med* 1987;**147**:2155–2161.
19 Apolone G, Bertolini G, D'Amico R, Iapichino G, Cattaneo A, De Salvo G *et al*. The performance of SAPS II in a cohort of patients admitted to 99 Italian ICUs: results from GiViTI. Gruppo Italiano per la Valutazione degli interventi in Terapia Intensiva. *Intensive Care Med* 1996;**22**:1368–1378.
20 Moreno R, Miranda DR, Fidler V, Van Schilfgaarde R. Evaluation of two outcome prediction models on an independent database. *Crit Care Med* 1998;**26**:50–61.
21 Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. *Intensive Care Med* 2003; **29**:249–256.
22 Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med* 1991;**10**:1213–1226.
23 Harrell Jr FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;**3**:143–152.
24 Beck DH, Smith GB, Pappachan JV. The effects of two methods for customising the original SAPS II model for intensive care patients from South England. *Anaesthesia* 2002; **57**:785–793.
25 Teres D, Lemeshow S. When to customize a severity model. *Intensive Care Med* 1999;**25**:140–142.

Accepted 19 June 2007
Available online 2 August 2007