Original article

Evaluation of POSSUM in patients with oesophageal cancer undergoing resection

K. D. Zafirellis, A. Fountoulakis, K. Dolan, S. P. L. Dexter, I. G. Martin and H. M. Sue-Ling

Division of Surgery, The General Infirmary at Leeds, Leeds LS1 3EX, UK
Correspondence to: Mr K. D. Zafirellis (e-mail: medkd@leeds.ac.uk)

Background: The Physiogical and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM) has been used to produce a numerical estimate of expected mortality and morbidity after a variety of general surgical procedures. The aim of this study was to evaluate the ability of POSSUM to predict mortality and morbidity in patients undergoing oesophagectomy.

Methods POSSUM predictor equations for morbidity and mortality were applied retrospectively to 204 patients who had undergone oesophagectomy for cancer. Observed morbidity and mortality rates were compared with rates predicted by POSSUM using the Hosmer–Lemeshow goodness-of-fit test. Evaluation of the discriminative capability of POSSUM predictor equations was performed using receiver–operator characteristic (ROC) curve analysis.

Results: The observed and predicted mortality rates were 12·7 and 19·1 per cent respectively, and the respective morbidity rates were 53·4 and 62·3 per cent. However, the POSSUM model showed a poor fit with the data both for the observed 30-day mortality (χ^2 = 16·26, P = 0·002) and morbidity (χ^2 = 63·14, P < 0·001) using the Hosmer–Lemeshow test. ROC curve analysis revealed that POSSUM had poor predictive accuracy both for mortality (area under curve 0·62) and morbidity (area under curve 0·55).

Conclusion These data suggest that POSSUM does not accurately predict mortality and morbidity in patients undergoing oesophagectomy and must be modified.

Paper accepted 29 April 2002 British Journal of Surgery 2002, **89,** 1150–1155

Introduction

Morbidity and mortality rates are crude outcome measures. Risk-adjusted models, taking into account variations in case mix, reveal more about the quality of care[1–5]. Acute Physiology And Chronic Health Evaluation (APACHE) II has found wide application in intensive care patients[2] and, although it produces a numerical estimate of mortality, it ignores morbidity rates and does not take into consideration the severity of the surgical insult. The Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM) is a risk prediction model based on 12 characteristics of the patient and six characteristics of the operation[5]. It is superior to APACHE II for the prediction of postoperative death in patients undergoing surgery in a high-dependency unit[6]. POSSUM has been used to make comparisons between different vascular[7,8] and colorectal[9] surgical units, and to compare individual surgeons' performance within a single unit[10,11]. However, POSSUM was developed for quality assessment in general surgical units and it would not be appropriate to use this model for specific subgroups of patients unless good model performance within those subgroups could be demonstrated. This is the first study to assess the accuracy of POSSUM in predicting mortality and morbidity in patients with oesophageal cancer undergoing resection.

Patients and methods

Between January 1990 and December 1999, 213 patients with oesophageal cancer underwent resection at The General Infirmary at Leeds. Selection of patients for surgery was based on surgeons' 'end of the bed' assessment, supported by pulmonary function tests, arterial blood gas measurement, electrocardiography and, more recently, echocardiography. Ninety-two per cent of operations were performed by four consultants with a special interest in upper gastrointestinal surgery. Nine patients were excluded from the study owing to incomplete data despite extensive tracking of case notes. The remaining 204 patients were scored retrospectively using POSSUM, and the predicted risk of morbidity and death was calculated for

each patient according to the following previously described logistic regression equations[5]:

$$\log_e[R_1/(1 - R_1)] = -5{\cdot}91 + (0{\cdot}16 \times \text{physiological score}) + (0{\cdot}19 \times \text{operative severity score})$$

where R_1 = risk of morbidity;

$$\log_e[R_2/(1 - R_2)] = -7{\cdot}04 + (0{\cdot}13 \times \text{physiological score}) + (0{\cdot}16 \times \text{operative severity score})$$

where R_2 = risk of death.

The definitions and classification of morbidity have been described previously[5] and mortality was determined at 30 days.

Statistical analysis

The performance of POSSUM in predicting mortality and morbidity was analysed by measures of calibration and discrimination[12]. Model calibration was assessed using the Hosmer–Lemeshow goodness-of-fit test and the corresponding calibration curves[13]. The patients were divided into risk groups on the basis of their predicted mortality and morbidity. The observed and predicted numbers of patients who experienced the event (death or complication) and those who did not were determined for each risk group. Summing the probabilities of mortality or morbidity for all patients in a risk group produced the predicted number of deaths or complications in that risk group. The discrepancies between the observed and predicted outcomes in these groups were tested using the χ^2 goodness-of-fit test. In this test $P > 0{\cdot}05$ indicates that the model is performing well.

Model discrimination was assessed using the area under the receiver–operator characteristic (ROC) curve (AUC)[14] to evaluate how well the model distinguished patients who experienced the event (death or complication) from those who did not. This statistic represents the concordance between predicted probabilities and observed outcomes for all possible pairs of patients with different outcome status. The AUC is used as an index of model discrimination; it ranges from 0·5 for chance performance to 1·0 for perfect prediction. In all analyses, confidence intervals were chosen at 95 per cent and $P < 0{\cdot}05$ was considered significant.

Results

Two hundred and four patients were scored; 146 (71·6 per cent) were men and the median age was 66 (range 29–89) years. More than three-quarters of patients underwent an Ivor–Lewis oesophagectomy; the remainder had a transhiatal or McKeown oesophagectomy. One-quarter of patients developed respiratory infection, 8 per cent had a wound infection, 3 per cent suffered a myocardial infarct and 1 per cent a pulmonary embolus. Gastrografin (Schering Health Care, Burgess Hill, UK) swallow was routinely performed on the seventh day after operation; an anastomotic leak was detected in less than 10 per cent. The median physiological score assigned by POSSUM was 16 (range 12–33) and the median operative severity score was 19 (range 14–34).

Validation of the POSSUM mortality equation

Twenty-six patients (12·7 per cent) died within 30 days following operation compared with a predictive value of 39 (19·1 per cent), giving a standardized mortality ratio of 0·66 (95 per cent confidence interval (c.i.) 0·43 to 0·97). However, the Hosmer–Lemeshow goodness-of-fit test indicated that the POSSUM mortality equation had a significant lack of fit with the data ($\chi^2 = 16{\cdot}26$, 4 d.f., $P = 0{\cdot}002$) (*Table 1*). The calibration curve for the POSSUM mortality equation applied to the data showed a discrepancy between actual and predicted mortality rates,

Table 1 Hosmer–Lemeshow goodness-of-fit test for the POSSUM mortality equation

Predicted risk of death (%)	No. of patients	No. of survivors		No. of deaths	
		Observed	Predicted	Observed	Predicted
> 0 to ⩽10	56	53	51	3	5
> 10 to ⩽20	83	70	71	13	12
> 20 to ⩽30	33	28	25	5	8
> 30 to ⩽40	14	11	9	3	5
> 40 to ⩽50	9	8	5	1	4
> 50 to ⩽100	9	8	4	1	5
> 0 to ⩽100	204	178	165	26	39

POSSUM, Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity. $\chi^2 = 16{\cdot}26$, 4 d.f., $P = 0{\cdot}002$

especially in the moderate- and high-risk groups (*Fig. 1*). ROC curve analysis revealed that POSSUM had a poor discriminatory capability for death with a ROC curve close to the diagonal line of chance (AUC = 0·62 (95 per cent c.i. 0·52 to 0·71)) (*Fig. 2*).

Validation of the POSSUM morbidity equation

Postoperative complications developed in 109 patients (53·4 per cent) compared with a predictive value of 127 (62·3 per cent), giving a standardized morbidity ratio of 0·86 (95 per cent c.i. 0·70 to 1·03). The Hosmer–Lemeshow goodness-of-fit test indicated that the POSSUM morbidity equation did not fit the data well (χ^2 = 63·14, 6 d.f., P < 0·001) (*Table 2*). The calibration curve showed a discrepancy between the actual and predicted morbidity rates (*Fig. 3*). ROC curve analysis revealed that POSSUM had poor discriminatory power for morbidity (AUC = 0·55 (95 per cent c.i. 0·47 to 0·63)) (*Fig. 4*).

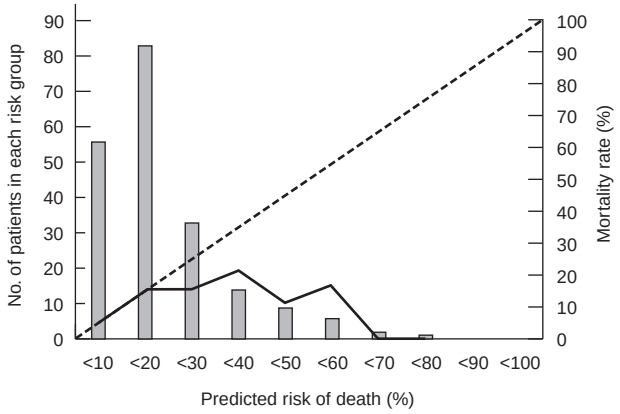

Fig. 1 Calibration curve for surgical mortality. The curve represents the proportion of patients dying within 30 days following operation according to their predicted risk of death estimated by the Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM) mortality equation. The dashed diagonal line represents the perfect predictive ability when observed and predicted mortality are equal

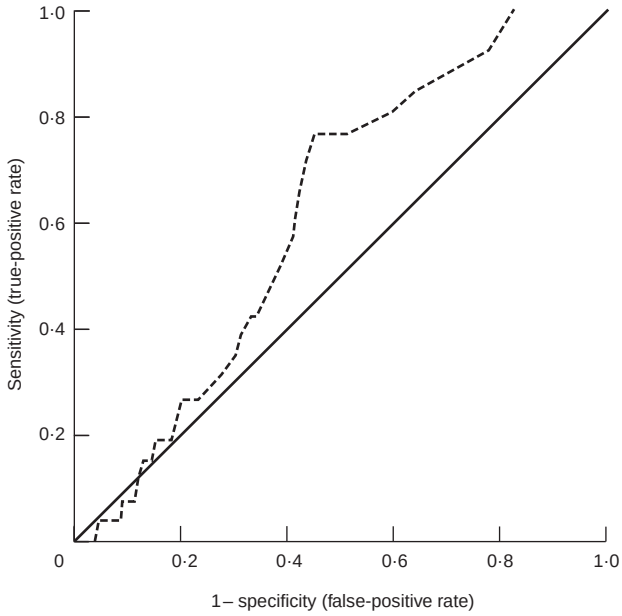

Fig. 2 Receiver–operator characteristic curve for mortality. The diagonal line represents predictive accuracy no better than chance. Area under the curve 0·62 (95 per cent confidence interval 0·52 to 0·71)

Table 2 Hosmer–Lemeshow goodness-of-fit test for the POSSUM morbidity equation

Predicted risk of morbidity (%)	No. of patients	No. of survivors		No. of deaths	
		Observed	Predicted	Observed	Predicted
> 0 to ⩽30	4	3	3	1	1
> 30 to ⩽40	26	11	17	15	9
> 40 to ⩽50	26	14	14	12	12
> 50 to ⩽60	47	28	21	19	26
> 60 to ⩽70	33	11	11	22	22
> 70 to ⩽80	28	12	7	16	21
> 80 to ⩽90	24	9	3	15	21
> 90 to ⩽100	16	7	1	9	15
> 0 to ⩽100	204	95	77	109	127

POSSUM, Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity. χ^2 = 63·14, 6 d.f., P < 0·001

Comparison of patients who died within 30 days with survivors

To investigate whether factors other than those recorded by POSSUM were significant in determining outcome, patients who died within 30 days of surgery were compared with those who survived longer. There were no significant differences in sex, tumour type, site and stage, type and completeness of resection, and use of neoadjuvant treatment between the two groups. Interestingly, there were no significant differences in POSSUM score for patients who died compared with survivors (*Table 3*).

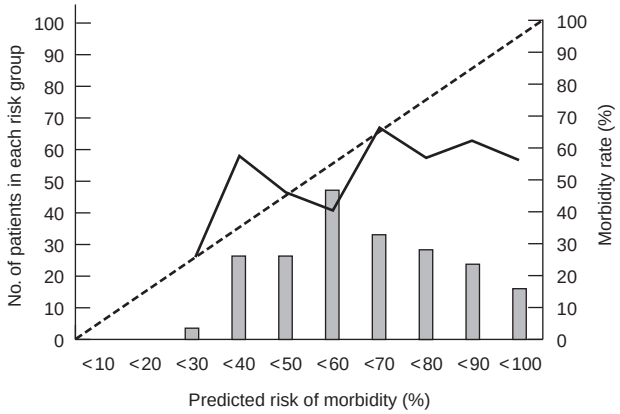

Fig. 3 Calibration curve for surgical morbidity. The curve represents the proportion of patients with complications following operation according to their predicted risk of morbidity estimated by the Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM) morbidity equation. The dashed diagonal line represents the perfect predictive ability when observed and predicted morbidity are equal

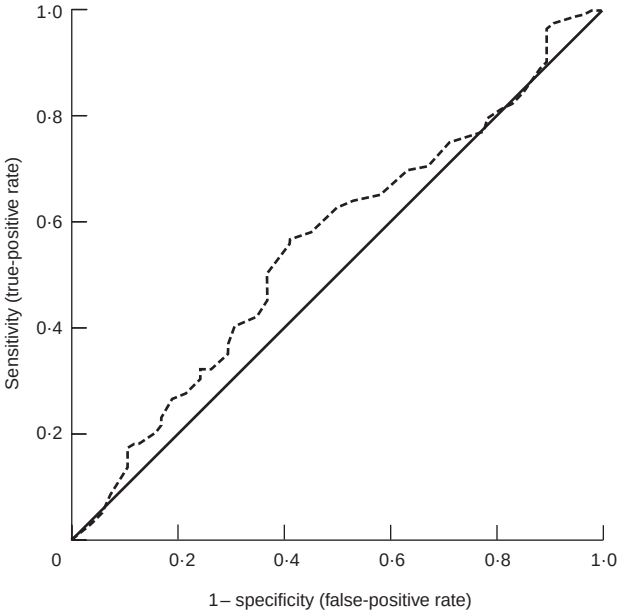

Fig. 4 Receiver–operator characteristic curve for morbidity. The diagonal line represents predictive accuracy no better than chance. Area under the curve 0·55 (95 per cent confidence interval 0·47 to 0·63)

Discussion

Before introduction to clinical practice, risk prediction models must be validated in a patient population independent from the population in which they were generated. Application of a model to a population with a different case

Table 3 Comparison of patients who died within 30 days and survivors

	Patients who died* (n = 26)	Survivors (n = 178)
Sex		
Male	19 (73·1)	127 (71·3)
Female	7 (26·9)	51 (28·7)
Tumour site		
Upper third	0 (0)	1 (0·6)
Middle third	7 (26·9)	38 (21·3)
Lower third	19 (73·1)	139 (78·1)
Histological type		
Adenocarcinoma	18 (69·2)	138 (77·5)
Squamous cell	8 (30·8)	37 (20·8)
Mixed	0 (0)	2 (1·1)
Oat cell	0 (0)	1 (0·6)
UICC stage		
I	3 (11·5)	25 (14·0)
IIA	8 (30·8)	39 (21·9)
IIB	3 (11·5)	18 (10·1)
III	9 (34·6)	67 (37·6)
IV	3 (11·5)	29 (16·3)
R category		
R0	14 (53·8)	85 (47·8)
R1	10 (38·5)	72 (40·4)
R2	2 (7·7)	21 (11·8)
Type of operation		
Ivor–Lewis	20 (76·9)	138 (77·5)
Transhiatal	0 (0)	9 (5·1)
McKeown	2 (7·7)	5 (2·8)
Thoracopically assisted	3 (11·5)	19 (10·7)
Left thoracoabdominal	1 (3·8)	7 (3·9)
Neoadjuvant treatment	4 (15·4)	35 (19·7)
POSSUM		
Mean (range) predicted mortality risk (%)	20·9 (9·2–52·1)	19·1 (4·3–72·2)
Mean (range) predicted morbidity risk (%)	69·2 (42·7–94·1)	61·4 (21·0–97·6)

Values in parentheses are percentages unless otherwise indicated. *Within 30 days. UICC, Union Internacional Contra la Cancrum; POSSUM, Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity. There were no significant differences between the two groups

mix from the sample from which the model was derived may produce misleading outcome predictions[15]. Model validation should be performed using measures of calibration and discrimination[12], as these are complementary and provide different information about model performance[16]. This is the first study to evaluate POSSUM in estimating mortality and morbidity using measures of both calibration and discrimination.

Calibration evaluates the degree of correspondence between the estimated probabilities of mortality and morbidity produced by the model and the actual experience of patients in various risk strata[12]. Using the Hosmer–Lemeshow test[13], POSSUM showed a poor fit with the data for both mortality and morbidity. Across the risk categories POSSUM performed relatively well in low-risk groups of patients and overpredicted the number of deaths and complications in high-risk groups. Overall, POSSUM overpredicted mortality and morbidity rates in patients with oesophageal cancer undergoing resection.

Model discrimination evaluates the ability of the model to distinguish patients who will experience the event of interest (death or complication) from those who will not[12]. ROC curve analysis revealed that POSSUM had poor discriminative capability for both mortality and morbidity in patients undergoing oesophagectomy for cancer.

The published literature on POSSUM has focused on the use of the model as a tool for comparative surgical audit[7–11,17,18], and few independent studies have attempted validation of POSSUM. In general surgery the Hosmer–Lemeshow test revealed that POSSUM overpredicted overall risk of death, performing worst among low-risk general surgical patients[19,20]. In an attempt to improve the predictive ability of the POSSUM mortality equation the authors proposed a simple adjustment of the model using the same variables in a different formula (Portsmouth POSSUM). In vascular surgery a 'linear' type of analysis found that POSSUM predicted morbidity well but significantly overpredicted mortality[21]. Although none of these studies validated its chosen model by measures of both calibration and discrimination, they agree with this study that POSSUM overpredicts death.

The original paper describing POSSUM evaluated its overall performance in the general surgical population without reporting the uniformity of fit of the model across the various surgical subspecialties[5]. The poor performance of POSSUM in this study may be a reflection that the model was originally based on a general surgical population that contained few if any patients undergoing oesophagectomy, and the score assigned to oesophagectomy may be too high. Amendment of the operative severity score may increase the predictive value of POSSUM but a larger data set is required for this. With respect to the physiological score of patients undergoing oesophagectomy, POSSUM does not include variables such as arterial blood gases, pulmonary function tests and echocardiography. The introduction of these variables into the physiological equation may improve its predictive value.

POSSUM cannot be used to audit oesophagectomy for cancer because of overprediction of mortality and morbidity. To develop new equations for patients with oesophageal cancer undergoing resection, a national database of oesophageal resections is required, with physiological scores potentially including arterial blood gases, pulmonary function tests and echocardiography. Logistic regression analysis would determine a POSSUM model for patients with oesophageal cancer undergoing oesophagectomy and this model must be validated in an independent population by measures of both calibration and discrimination.

To describe surgical mortality and morbidity in a meaningful way, it is essential that an accurate method of comparing mortality and morbidity rates following oesophagectomy is developed as soon as possible.

References

1 Goldman L, Caldera DL, Nussbaum SR, Southwick FS, Krogstad D, Murray B *et al.* Multifactorial index of cardiac risk in noncardiac surgical procedures. *N Engl J Med* 1977; **297:** 845–50.
2 Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; **13:** 818–29.
3 Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D *et al.* A simplified acute physiology score for ICU patients. *Crit Care Med* 1984; **12:** 975–7.
4 Lemeshow S, Teres D, Pastides H, Avrunin JS, Steingrub JS. A method for predicting survival and mortality of ICU patients using objectively derived weights. *Crit Care Med* 1985; **13:** 519–25.
5 Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. *Br J Surg* 1991; **78:** 355–60.
6 Jones DR, Copeland GP, de Cossart L. Comparison of POSSUM with APACHE II for prediction of outcome from a surgical high-dependency unit. *Br J Surg* 1992; **79:** 1293–6.
7 Copeland GP, Jones D, Wilcox A, Harris PL. Comparative vascular audit using the POSSUM scoring system. *Ann R Coll Surg Engl* 1993; **75:** 175–7.
8 Treharne GD, Thompson MM, Whiteley MS, Bell PR. Physiological comparison of open and endovascular aneurysm repair. *Br J Surg* 1999; **86:** 760–4.
9 Sagar PM, Hartley MN, Mancey-Jones B, Sedman PC, May J, MacFie J. Comparative audit of colorectal resection with the POSSUM scoring system. *Br J Surg* 1994; **81:** 1492–4.
10 Copeland GP, Sagar P, Brennan J, Roberts G, Ward J, Cornford P *et al.* Risk-adjusted analysis of surgeon performance: a 1-year study. *Br J Surg* 1995; **82:** 408–11.

11 Sagar PM, Hartley MN, MacFie J, Taylor BA, Copeland GP. Comparison of individual surgeon's performance. Risk-adjusted analysis with POSSUM scoring system. *Dis Colon Rectum* 1996; **39:** 654–8.
12 Lemeshow S, Le Gall JR. Modeling the severity of illness of ICU patients. A systems update. *JAMA* 1994; **272:** 1049–55.
13 Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol* 1982; **115:** 92–106.
14 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143:** 29–36.
15 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:** 2155–61.
16 Hadorn DC, Keeler EB, Rogers WH, Brook RH. *Assessing the Performance of Mortality Prediction Models*. Santa Monica, California: RAND Corporation, 1993.
17 Curran JE, Grounds RM. Ward *versus* intensive care management of high-risk surgical patients. *Br J Surg* 1998; **85:** 956–61.
18 Jones HJ, Coggins R, Lafuente J, de Cossart L. Value of a surgical high-dependency unit. *Br J Surg* 1999; **86:** 1578–82.
19 Whiteley MS, Prytherch DR, Higgins B, Weaver PC, Prout WG. An evaluation of the POSSUM surgical scoring system. *Br J Surg* 1996; **83:** 812–15.
20 Prytherch DR, Whiteley MS, Higgins B, Weaver PC, Prout WG, Powell SJ. POSSUM and Portsmouth POSSUM for predicting mortality. Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity. *Br J Surg* 1998; **85:** 1217–20.
21 Midwinter MJ, Tytherleigh M, Ashley S. Estimation of mortality and morbidity risk in vascular surgery using POSSUM and the Portsmouth predictor equation. *Br J Surg* 1999; **86:** 471–4.

