Validation of the Rockall risk scoring system in upper gastrointestinal bleeding

E M Vreeburg, C B Terwee, P Snel, E A J Rauws, J F W M Bartelsman, J H P vd Meulen, G N J Tytgat

Abstract

***Background*—Several scoring systems have been developed to predict the risk of rebleeding or death in patients with upper gastrointestinal bleeding (UGIB). These risk scoring systems have not been validated in a new patient population outside the clinical context of the original study.**

***Aims*—To assess internal and external validity of a simple risk scoring system recently developed by Rockall and coworkers.**

***Methods*—Calibration and discrimination were assessed as measures of validity of the scoring system. Internal validity was assessed using an independent, but similar patient sample studied by Rockall and coworkers, after developing the scoring system (Rockall's validation sample). External validity was assessed using patients admitted to several hospitals in Amsterdam (Vreeburg's validation sample). Calibration was evaluated by a χ^2 goodness of fit test, and discrimination was evaluated by calculating the area under the receiver operating characteristic (ROC) curve.**

***Results*—Calibration indicated a poor fit in both validation samples for the prediction of rebleeding ($p<0.0001$, Vreeburg; $p=0.007$, Rockall), but a better fit for the prediction of mortality in both validation samples ($p=0.2$, Vreeburg; $p=0.3$, Rockall). The areas under the ROC curves were rather low in both validation samples for the prediction of rebleeding (0.61, Vreeburg; 0.70, Rockall), but higher for the prediction of mortality (0.73, Vreeburg; 0.81, Rockall).**

***Conclusions*—The risk scoring system developed by Rockall and coworkers is a clinically useful scoring system for stratifying patients with acute UGIB into high and low risk categories for mortality. For the prediction of rebleeding, however, the performance of this scoring system was unsatisfactory.**

(*Gut* 1999;**44**:331–335)

Keywords: upper gastrointestinal bleeding; risk scoring; prognostic factors; rebleeding; mortality

Department of Gastroenterology and Hepatology, Academic Medical Centre, Amsterdam, The Netherlands
E M Vreeburg
E A J Rauws
J F W M Bartelsman
G N J Tytgat

Department of Gastroenterology, Slotervaart Hospital, Slotervaart, The Netherlands
P Snel

Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, Amsterdam, The Netherlands
C B Terwee
J H P vd Meulen

Correspondence to:
Ms C B Terwee, Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands.

Accepted for publication
9 October 1998

Upper gastrointestinal bleeding (UGIB) represents a common emergency in clinical practice with an incidence of 50–150 per 100 000 people per year.[1–5] The mortality rate of UGIB decreased slightly following the introduction of endoscopic intervention modalities in the 1980s and better care in high dependency bleeding units,[1 6 7] but still varies between 4% and 14%.[8–11] Rebleeding is considered the most important risk factor for mortality and occurs in 10–30% of those successfully treated.[12]

Different clinical and endoscopic factors associated with an increased risk for rebleeding and mortality after admission for UGIB have been described,[2 8–10] although there is still considerable disagreement about what the most important prognostic factors are. Several risk scoring systems have been proposed to classify patients into high and low risk groups for rebleeding or mortality based on multivariate analyses.[12–18] These scoring systems can be used to select low risk patients for early discharge or outpatient treatment, and to select high risk patients for intensive care treatment, which improves efficiency of current therapy. Furthermore, risk scoring systems can be used to stratify patients who are included in clinical trials which study the effectiveness of endoscopic or other medical interventions. Unfortunately, the complexity and variability of these scoring systems limits their application in routine clinical practice.

More importantly however, the performance of most of these scoring systems has never been validated in a population of new patients. Validation refers to calibration, or the amount of agreement between predicted probabilities and observed percentages of rebleeders/deaths in different risk groups, and discrimination, or the ability of a scoring system to distinguish patients who rebleed/die from patients who do not rebleed/live.[19 20] Validity can be separated into internal and external validity: internal validity indicates whether the results of the analysis hold in future patients who are included according to the same criteria and within the same clinical context as the patients in the original study; external validity or generalisability refers to the performance of the scoring system in patients outside the study context, for example, patients in other hospitals. External validity is especially important when scoring systems are used to predict outcome in daily practice, because it is well known that scoring systems (or models in general) perform less well in patient samples outside the clinical context in which these models are developed.[19]

Therefore, the aim of this study was to assess the external validity of a scoring system for predicting rebleeding and death after admis-

Abbreviations used in this paper: ROC curve, receiver operating characteristic curve; SRH, stigmata of recent haemorrhage; UGIB, upper gastrointestinal bleeding.

Table 1 The Rockall risk scoring system

	Score			
Variable	*0*	*1*	*2*	*3*
Age (years)	<60	60–79	⩾80	
Shock	"No shock": pulse <100 + systolic BP⩾100 mm Hg	"Tachycardia": pulse ⩾100 + systolic BP ⩾100 mm Hg	"Hypotension": systolic BP ⩾100 mm Hg	
Comorbidity	No major comorbidity		Cardiac failure, ischaemic heart disease, any major comorbidity	Renal failure, liver failure, disseminated malignancy
Diagnosis	Mallory Weiss tear, no lesion identified and no SRH/blood	All other diagnoses	Malignancy of upper GI tract	
Major SRH	None or dark spot only		Blood in upper GI tract, adherent clot, visible or spurting vessel	
"Translation" of our comorbidity scale				
Comorbidity	No or mild coexisting illnesses (e.g. ECG abnormalities without symptoms)	Moderate coexisting illnesses (e.g. hypertension stable with medication)	Severe coexisting illnesses (diseases which need immediate treatment: e.g. cardiac failure)	Life threatening diseases (e.g. end stage malignancies, renal failure)

Major SRH, major stigmata of recent haemorrhage (active bleeding or visible vessel); GI, gastrointestinal; BP, blood pressure.

sion for UGIB, that was recently developed by Rockall *et al*,[13] in order to investigate its performance in a Dutch patient population. Rockall *et al* included 4185 cases of acute UGIB from 74 hospitals in the UK over a four month period in 1993. Their scoring system was based on multivariate analysis of information from history, examination, blood tests, and endoscopic investigation. We used Rockall's risk scoring system to classify patients, admitted to several hospitals in Amsterdam, into different risk groups. Although the scoring system was originally developed to predict mortality, Rockall *et al* suggested in their article that it could also be used for the prediction of rebleeding. We therefore applied the same scoring system to predict rebleeding as well.

We also assessed the internal validity of the risk scoring system by using a second patient group of Rockall *et al*, included according to the same criteria and within the same clinical context as the patients in the original study, and whose prognostic scores and outcomes are presented in the original paper of Rockall *et al* (audit 2, presented in table V (A, B)[13]). We refer to this patient group as Rockall's validation sample.

We assessed the calibration of the risk scoring system by a goodness of fit test and the discriminative ability by receiver operating characteristic (ROC) analysis.[21]

Patients and methods

PATIENTS

We prospectively studied all patients who were consecutively admitted to the endoscopy ward of two university and 10 regional hospitals in the same Amsterdam area (catchment area, 1 610 900 persons) with symptoms of haematemesis, melaena, haematochezia, or blood admixture on nasogastric aspiration who were suspected of having acute UGIB.[11] Patients were included in the study from July 1993 until July 1994. Patients who developed an acute upper gastrointestinal bleed while being hospitalised for other diseases were also included in the study.

Data were collected using a standard protocol and included: demographic characteristics, symptoms and signs of the gastrointestinal bleeding episode, symptoms and history of peptic ulcer and/or liver disease, coexisting illnesses, drug history, laboratory results, endoscopic intervention, medical treatment, transfusion requirements, rebleeding incidence, surgical treatment, complications, duration of hospitalisation, cause of death, and stigmata of recent haemorrhage (SRH), which included spurting arterial bleeding, oozing of blood, non-bleeding visible vessel, overlying clot, or haematin covered ulcer base.

Coexisting illnesses were classified according to the ICED scale (Index of Co-Existing Disease, National Auxiliary Publication Service[22]). This scale classifies diseases other than the gastrointestinal bleeding disorder into mild, moderate, severe, or life threatening diseases, based on a series of preset definitions.

Rebleeding was defined as a new episode of bleeding during hospitalisation, after the initial bleeding had stopped, that manifested as

Table 2 Distribution of patients in the risk score groups, calculated with the Rockall risk score, for the Rockall validation sample and for our own patient group

	*Predicted probabilities**		*Rockall's validation sample*			*Vreeburg's validation sample*		
Risk score	*Rebleeding (%)*	*Mortality (%)*	*Number of patients*	*Rebleeding (%)*	*Mortality (%)*	*Number of patients*	*Rebleeding (%)*	*Mortality (%)*
0	4.9	0	48	4.2	0	11	9.1	0
1	3.4	0	131	4.6	0	36	3.8	0
2	5.3	0.2	142	7.7	0	71	8.5	1.4
3	11.2	2.9	162	11.7	1.8	145	13.8	7.6
4	14.1	5.3	176	15.3	8.0	175	11.4	9.7
5	24.1	10.8	199	24.6	10.6	178	16.3	10.7
6	32.9	17.3	137	27.0	11.7	142	22.5	17.6
7	43.8	27.0	96	40.6	25.0	107	20.6	24.3
8+	41.8	41.1	89	37.1	40.4	86	26.7	46.5
Total	18.9	10.0	1180	18.9	9.7	951	16.4	13.9

*Predicted probabilities based on observed percentages in original patient sample (Rockall, table V(B)[13]).

recurrent haematemesis, haematochezia, fresh blood in the nasogastric aspirate, or circulatory instability. Further haemorrhage, neccessitating surgery, was also defined as rebleeding. Mortality was defined as death within the hospitalisation period.

We refer to this patient group as Vreeburg's validation sample.

APPLICATION OF THE RISK SCORING SYSTEM

Table 1 shows the risk scoring system developed by Rockall *et al.*[13] The scoring system represents a simplified summary of the results of a logistic regression analysis and includes three clinical variables (age, shock, and comorbidity) and two endoscopic variables (diagnosis and major SRH), each categorised and scored with 0–3 points, to give a maximum score of 11 points. We used this scoring system to assess the individual risk score for each patient in Vreeburg's validation sample. Risk scores for Rockall's validation sample were obtained from table V(B) of Rockall's paper.[13] Scores of $\leqslant 3$ and scores of $\geqslant 8$ were taken together as one category because of the low numbers in each of these outcome categories. We used the observed percentages of rebleeders/deaths in each risk category in the original patient sample of Rockall (presented in table IV(B) in Rockall's paper[13]) as the predicted probabilities of rebleeding/mortality for both validation samples.

Because our classification of coexisting illnesses according to the ICED scale did not completely correspond to the classification of comorbidity used by Rockall, we scored the ICED classification as follows: none or mild coexisting illnesses received zero points, moderate illnesses received one point, severe illnesses received two points, and life threatening conditions received three points (table 1) (Rockall, personal communication).

VALIDATION OF THE RISK SCORING SYSTEM

Internal and external validity of the risk scoring system was assessed using a χ^2 goodness of fit test as a measure of (model) calibration and the area under the ROC curve as a measure of (model) discrimination. The goodness of fit test evaluates the degree of correspondence between predicted probabilities and observed percentages of rebleeders/deaths. If the observed percentages of rebleeders/deaths are close to the predicted probabilities, the risk scoring system is considered to be well calibrated.[20] The area under the ROC curve evaluates the ability of the risk scoring system to distinguish patients who rebled/died from those who did not. In the ROC curve, pairs of true positive and false positive rates are plotted, based on 2×2 classification tables of predicted and observed rebleeding/mortality, that can be constructed for each risk probability cut off point.[18] The area under the curve (AUC) is a measure of the discriminative value of the risk scoring system. If this area is 0.50, the scoring system is performing no better than the toss of a coin. An area of 1.0 would reflect a perfect discriminative ability.[23]

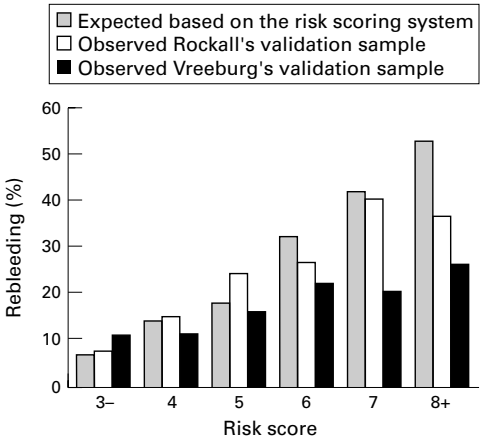

Figure 1 Expected versus observed rebleeding by risk score.

Results

In total, 951 patients were included, with a median age of 71 years (range 2–100), of whom 25% were older than 80 years, and 60% were men. The rate of rebleeding was 16% (n=156) and the mortality rate was 14% (n=132) during hospitalisation. Detailed characteristics of the study population are described elsewhere.[11]

The distribution of patients over the risk categories, as determined by the risk scoring system of Rockall, and the observed percentages of rebleeding and mortality in each risk category is shown in table 2 for Rockall's validation sample (from table V(B) of Rockall[13]) and for Vreeburg's validation sample.

VALIDITY OF THE RISK SCORING SYSTEM

Calibration

Figure 1 shows the predicted probabilities of rebleeding based on the original patient sample of Rockall compared with the observed rebleeding percentages of Rockall's validation sample and Vreeburg's validation sample. In the lowest risk categories, the predicted probabilities were lower than the observed rebleeding rate, while in the highest risk categories the predicted probabilities were higher than the observed rebleeding rate. The goodness of fit test indicates a lack of fit for

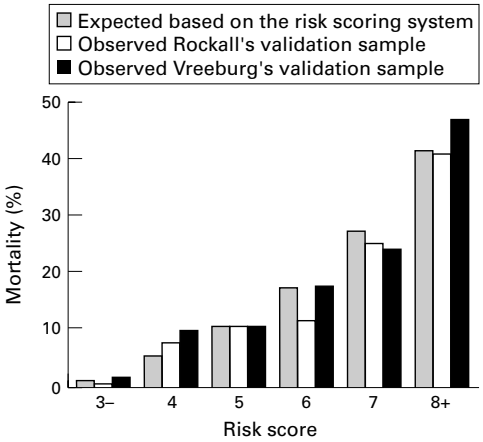

Figure 2 Expected versus observed mortality by risk score.

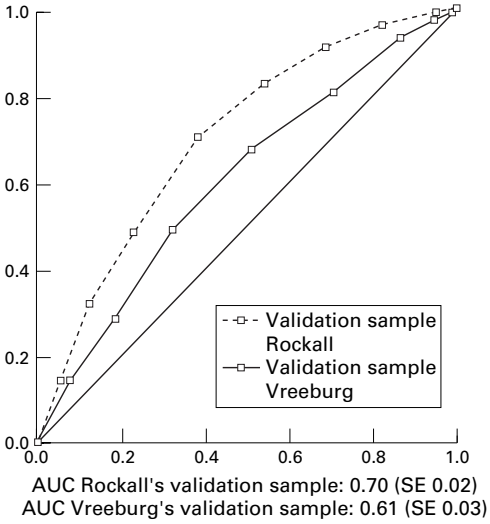

Figure 3 Discriminative ability of the Rockall scoring system for the prediction of rebleeding, expressed as AUC for the internal validation sample (Rockall) and the external validation sample (Vreeburg).

both validation samples (χ^2=17.6, df=6, p=0.007 for Rockall's validation sample, and χ^2=61.6, df=6, p<0.0001 for Vreeburg's validation sample).

In fig 2 the corresponding findings for the prediction of mortality are shown. Here the correspondence between predicted and observed rates was better for both validation samples (χ^2=7.08, df=6, p=0.3 for Rockall's validation sample, and χ^2=9.3, df=6, p=0.2 for Vreeburg's validation sample) indicating a better fit.

Overall, the predicted probabilities for rebleeding and mortality were closer to the observed rebleeding/mortality percentages of Rockall's validation sample than of Vreeburg's validation sample.

Discrimination

The discriminative abilities of the risk scoring system for the prediction of rebleeding and mortality are given in figs 3 and 4 respectively.

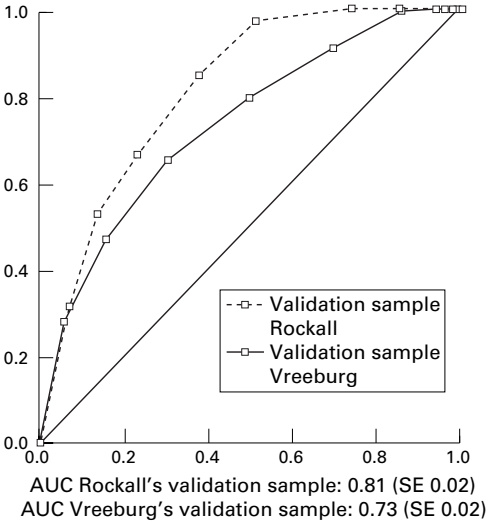

Figure 4 Discriminative ability of the Rockall scoring system for the prediction of mortality, expressed as AUC for the internal validation sample (Rockall) and the external validation sample (Vreeburg).

For rebleeding, the AUCs were 0.70 (SE 0.02) for Rockall's validation sample and 0.61 (SE 0.03) for Vreeburg's validation sample. For mortality, the AUCs were 0.81 (SE 0.02) for Rockall's validation sample and 0.73 (SE 0.02) for Vreeburg's validation sample.

As with calibration, the discriminative ability of the scoring system was better for the prediction of mortality than for the prediction of rebleeding. Furthermore, the discriminative ability, for rebleeding as well as for mortality, was better for Rockall's validation sample than for Vreeburg's validation sample.

Discussion

Overall, the internal and external validity of the risk scoring system, as assessed by calibration and discrimination, could be considered satisfactory for the prediction of mortality but not for the prediction of rebleeding. For the prediction of rebleeding, we observed a lack of fit for both validation samples, and the AUCs were rather low (0.70 and 0.61). For the prediction of mortality, we observed a better fit and higher AUCs (0.81 and 0.73). As expected, the internal validity was higher than the external validity.

It is important that the performance of such a risk scoring system is shown in a sample of new patients outside the original study context, especially when a scoring system is used to predict outcome for future patients, because it is well recognised that a scoring system tends to perform better in the population in which it is developed. Although Rockall's validation sample included new patients, who were not used in the development of the scoring system, these patients were included in the same hospitals using the same study protocol, and can be considered as the same type of patients. Therefore, the results in this patient sample indicate only the internal validation of the scoring system.

Our patient sample was slightly different from Rockall's original patient sample because our classification of comorbidity according to the ICED scale was different. However, we tried to adjust as accurately as possible the ICED classification to the classification of Rockall. Secondly, and probably more important, we used hospital mortality while Rockall used 30 day mortality. This implies that we included patients who died after 30 days while hospitalised. For these patients, the prediction of mortality might be more difficult and the scoring system of Rockall might not be applicable for these patients. This might have led to an underestimation of the validity of the scoring system. On the other hand, we might have missed patients who died within 30 days but after discharge from the hospital. However, we assume that this latter group of patients will be rather small and will therefore not influence the results of the study.

The disappointing performance of the risk scoring system in the prediction of rebleeding might partly be explained by the fact that the risk scoring system was originally developed for the prediction of mortality and not for the prediction of rebleeding. Possibly, other risk factors are more important for the prediction

of rebleeding than for the prediction of mortality, or the risk factors should be weighted differently. Other scoring systems, specifically developed for the prediction of rebleeding, such as the Baylor bleeding score,[24] are promising. Saeed *et al*[12] applied this scoring system to an external patient group who presented with major ulcer haemorrhage, and found higher rates of rebleeding in high risk patients, compared with low risk patients. However, formal calibration or discrimination of this scoring system has not yet been assessed.

The aim of this study was to validate a simple risk scoring system proposed by Rockall *et al*. We did not assess the performance of the original logistic regression model, from which the scoring system was derived, in the prediction of rebleeding and mortality. An inadequate translation of the model into the risk scoring system could lead to a bad performance of the scoring system. For example, it was unclear to us why Rockall *et al* included rebleeding as a variable in the logistic regression model, but did not include rebleeding in the risk scoring system. This might have influenced the weighting of the variables in the scoring system. However, we agree that rebleeding should not be included in a scoring system, because at the time of admission with UGIB, rebleeding is an outcome event instead of a prognostic variable.

We showed that the Rockall risk scoring system has unsatisfactory validity for the prediction of rebleeding in patients admitted with acute UGIB. However, the system appears to be useful for the stratification of patients into high and low risk groups for mortality. In the highest risk category (at least eight points), mortality in our patient sample was 46.5% (40/86). This group of patients will probably benefit most from early intensive care treatment. In the lowest risk category (three points or less), mortality was still 4.2% (11/263) but in patients with two points or less, mortality was only 0.8% (1/118). The mortality rate of patients with two points or less in a recent study of Jones *et al* also appeared to be low (1%).[25] Therefore, one could consider the selection of patients with two points or less for early discharge, after the bleeding had settled. Therefore, for one fifth of the patients adequate management of care could be given. However, in the relatively large intermediate group, in which mortality varied from 9.7% to 24.3%, patient management is less clear and better discrimination is necessary. Part of this group might be treated optimally in specialised medium care units, especially for the first 72 hours, which has been recognised as the period in which rebleeding usually develops. However, more than 75% of these patients did not die and should be selected for early discharge. For these patients, the risk scoring system could possibly be improved by adding additional prognostic variables such as prior H_2 receptor antagonist therapy, smoking, or liver and renal function disturbances. These factors were found to be additional significant predictors of mortality in a multivariate logistic regression analysis based on our patient sample (data not shown). Better discrimination of high and low risk patients in this intermediate group would lead to better patient care and cost effective management.

We conclude that the risk scoring system developed by Rockall *et al* is a clinically useful scoring system for stratifying patients with acute UGIB into high and low risk categories for mortality, and in patient samples outside the original study context, but probably could be improved for the intermediate risk category. For the prediction of rebleeding, however, the performance of this scoring system was unsatisfactory.

1 Johnston SJ, Jones PF, Kyle J, *et al*. Epidemiology and course of gastrointestinal haemorrhage in north-east Scotland. *BMJ* 1973;**3**:655–60.
2 Schiller KF, Truelove SC, Williams DG. Haematemesis and melaena, with special reference to factors influencing the outcome. *BMJ* 1970;**2**:7–14.
3 Cutler JA, Mendeloff AI. Upper gastrointestinal bleeding. Nature and magnitude of the problem in the US. *Dig Dis Sci* 1981;**26**(suppl 7):90–6S.
4 Wara P. Endoscopic management of the bleeding ulcer. A survey. *Dan Med Bull* 1986;**33**:1–11.
5 Yavorski RT, Wong RK, Maydonovitch C, *et al*. Analysis of 3,294 cases of upper gastrointestinal bleeding in military medical facilities. *Am J Gastroenterol* 1995;**90**:568–73.
6 La Vecchia C, Lucchini F, Negri E, *et al*. The impact of therapeutic improvements in reducing peptic ulcer mortality in Europe. *Int J Epidemiol* 1993;**22**:96–106.
7 Cook DJ, Guyatt GH, Salena BJ, *et al*. Endoscopic therapy for acute nonvariceal upper gastrointestinal hemorrhage: a meta-analysis. *Gastroenterology* 1992;**102**:139–48.
8 Morgan AG, Clamp SE. OMGE international upper gastrointestinal bleeding survey, 1978–1986. *Scand J Gastroenterol* 1988;**144**(suppl):51–8.
9 Katschinski B, Logan R, Davies J, *et al*. Prognostic factors in upper gastrointestinal bleeding. *Dig Dis Sci* 1994;**39**:706–12.
10 Branicki FJ, Coleman SY, Fok PJ, *et al*. Bleeding peptic ulcer: a prospective evaluation of risk factors for rebleeding and mortality. *World J Surg* 1990;**14**:262–70.
11 Vreeburg EM, Snel P, Bruijne JW, *et al*. Acute upper gastrointestinal bleeding in the Amsterdam area; incidence, diagnosis and clinical outcome. *Am J Gastroenterol* 1997;**92**: 236–43.
12 Saeed ZA, Ramirez FC, Hepps KS, *et al*. Prospective validation of the Baylor bleeding score for predicting the likelihood of rebleeding after endoscopic hemostasis of peptic ulcers. *Gastrointest Endosc* 1995;**41**:561–5.
13 Rockall TA, Logan RFA, Devlin HB, *et al*. Risk assessment after acute upper gastrointestinal haemorrhage. *Gut* 1996;**38**:316–21.
14 Bordley DR, Mushlin AI, Dolan JG, *et al*. Early clinical signs identify low-risk patients with acute upper gastrointestinal hemorrhage. *JAMA* 1985;**253**:3282–5.
15 Schein M, Gecelter G. APACHE II score in massive upper gastrointestinal haemorrhge from peptic ulcer: prognostic value and potential clinical applications. *Br J Surg* 1989;**76**: 733–6.
16 Pimpl W, Boeckl O, Waclawiczek HW, *et al*. Estimation of the mortality rate of patients with severe gastroduodenal hemorrhage with the aid of a new scoring system. *Endoscopy* 1987;**19**:101–6.
17 Clason AE, Macleod DAD, Elton RA. Clinical factors in the prediction of further hemorrhage or mortality in acute upper gastrointestinal haemorrhage. *Br J Surg* 1986;**73**: 985–7.
18 Morgan AG, McAdam WA, Walmsley GL, *et al*. Clinical findings, early endoscopy, and multivariate analysis in patients bleeding from the upper gastrointestinal tract. *BMJ* 1977;**2**:237–40.
19 Lemeshow S, Le Gall J. Modelling the severity of illness of ICU patients. *JAMA* 1994;**272**:1049–55.
20 Braitman LE, Davidoff F. Predicting clinical states in individual patients. *Ann Intern Med* 1996;**125**:406–12.
21 Breslow NE, Day NE, Davis W. *Statistical methods in cancer research. Vol. I—The analysis of case control studies*. Lyon: IARC Scientific Publications, no. 32, 1980.
22 Greenfield S, Apolone G, McNeil BJ, *et al*. The importance of co-existing disease in the occurence of postoperative complications and one-year recovery in patients undergoing total hip-replacement. *Med Care* 1993;**31**:141–54.
23 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29–36.
24 Saeed ZA, Winchester CB, Michaletz PA, *et al*. A scoring system to predict rebleeding from peptic ulcer: prognostic value and clinical applications. *Am J Gastroenterol* 1993;**88**: 1842–9.
25 Jones SGW, Davies R, Epworth I, *et al*. Use of the Rockall score to assess the effectivenesss of an upper gastrointestinal bleeding unit [abstract]. *Gut* 1996;**39**(suppl): A4.