Anaesthesia, 2005, **60**, pages 323–331

Assessing the applicability of scoring systems for predicting postoperative nausea and vomiting

J. E. van den Bosch,[1] C. J. Kalkman,[2] Y. Vergouwe,[3] W. A. Van Klei,[4] G. J. Bonsel,[5] D. E. Grobbee[6] and K. G. M. Moons[7]

1 PhD Student, 2 Professor, 4 Resident in Anaesthesiology, and 7 Associate Professor, Department of Perioperative Care and Emergency Medicine, University Medical Centre Utrecht, the Netherlands
1 PhD Student, 3 Assistant Professor, 6 Professor and 7 Associate Professor, Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, the Netherlands
5 Professor, Department of Public Health, University Medical Centre Amsterdam, Amsterdam, the Netherlands

Summary

We have validated two scoring systems for predicting postoperative nausea and vomiting, derived by Apfel *et al.* and Koivuranta *et al.* from 1388 adult inpatients undergoing a wide range of surgical procedures. The predictive accuracy of the scoring systems was evaluated in terms of the ability to discriminate between patients with and without postoperative nausea and vomiting (discrimination) and agreement between observed and predicted outcomes (calibration). Discrimination and calibration were less than expected based on previous reports, with both scoring systems providing risk predictions that were too extreme. The area under the ROC curve was 0.63 for Apfel *et al.*'s scoring system and 0.66 for Koivuranta *et al.*'s scoring system. Neither of the scoring systems provided a risk threshold for administering anti-emetic prophylaxis that yielded satisfying results in terms of predictive values, sensitivity and specificity. Hence, in their original forms, the scoring systems do not guarantee accurate prediction of the risk of postoperative nausea and vomiting in other patient populations. Koivuranta *et al.*'s scoring system appears to be more robust across different populations.

Correspondence to: Prof. Karel G. M. Moons
E-mail: k.g.m.moons@umcutrecht.nl
Accepted: 22 December 2004

Despite modern anaesthetic and surgical techniques, the incidence of postoperative nausea and vomiting (PONV) is still about 25–30% [1, 2]. Since PONV may result in extreme discomfort, it is unacceptable to neglect the prophylactic potential of anti-emetics to reduce the risk. However, routine administration of anti-emetics to all surgical patients is contra-indicated because of potential side-effects and costs [2, 3]. Instead, prophylactic anti-emetics should be selectively administered to patients at high risk of PONV, for example those identified before surgery using a scoring system for predicting PONV.

Several scoring systems have been proposed for predicting PONV within 24 h after surgery [4–7]. The best known scoring systems are those developed by Apfel *et al.* (Germany) [4] and Koivuranta *et al.* (Finland) [5]. Both scoring systems have shown good predictive accuracy in the datasets from which they were developed. Consequently, Apfel *et al.*'s scoring system was recently used to select high-risk patients for participation in a randomised trial that evaluated six interventions to reduce PONV [8]. Furthermore, several groups have already implemented Apfel *et al.*'s scoring system in daily practice, for example, to investigate whether rule-based prophylaxis reduced the overall incidence of PONV [9–11]. However, before widespread implementation in clinical practice, a scoring system should be validated in various clinical settings, as it is known that a scoring system generally predicts less accurately in new patients and in other settings than in patients and settings from which it was derived [12–15]. The more diverse the settings in which a scoring system has been tested and found to be accurate, the more likely it is that the scoring system will generalise to new settings. Despite this, the predictive accuracy of the two scoring systems for predicting PONV has hardly been tested in patients other than the populations from which the systems were developed [11, 16, 17]. Moreover, one of these validation studies [16] was performed in the institution where the

original scoring system was derived and another [11] in a small number of (predominantly female) patients undergoing a limited number of surgical procedures. We therefore tested the generalisability of Apfel *et al.*'s and Koivuranta *et al.*'s scoring systems to a large number of patients undergoing a wide range of surgical procedures.

Methods

Scoring systems and derivation datasets

Apfel et al.'s scoring system

Apfel *et al.*'s scoring system was derived from a combined data set of 1040 adult surgical inpatients aged ≥ 18 years, in two centres (Oulu, Finland and Wuerzburg, Germany) [4]. Postoperative nausea and vomiting was defined as at least one episode of nausea and/or vomiting within the first 24 h after surgery. Nausea was assessed on a binary scale at 2 h and on a 11-point numeric scale (0–10) at 24 h after surgery. The number of vomiting episodes was recorded for both intervals. The average incidence of PONV in the combined dataset was 44%. Descriptives for the combined dataset were not presented in the original paper, but could be estimated by averaging the numbers presented for the Oulu and the Wuerzburg datasets (Table 1, column 2). The formula of the original scoring system, developed using multivariable logistic regression analysis, was:

$$\begin{aligned} \textit{Risk of PONV} = 1/(1+\exp(&-2.28+1.27\times\textit{female gender}\\ &+0.65\times\textit{history of PONV or motion sickness}\\ &+0.72\times\textit{non-smoking}\\ &+0.78\times\textit{postoperative opioid use})). \end{aligned} \quad (1)$$

The expression '$-2.28 + 1.27 \times$ *female gender* $+ 0.65 \times$ *history of PONV or motion sickness* $+ 0.72 \times$ *non-smoking* $+ 0.78 \times$ *postoperative opioid use*' is called the linear predictor. In this expression, −2.28 represents the scoring system's intercept and the other numbers represent the regression coefficients (weights) of each corresponding predictor. The formula $1/(1 + \exp(-2.28 + \ldots))$ represents patients' baseline risk, which is increased by the presence of any of the other factors.

This scoring system was simplified to a four-item risk score, which was defined as the number of predictors present. If none, one, two, three or four of these predictors are present, the predicted risk of PONV is 10%, 21%, 39%, 61% and 79%, respectively. The original scoring systems and simplified risk scores (from Oulu and Wuerzburg) were internally validated by cross-validation (developed in one centre and validated in the other centre).

Koivuranta et al.'s scoring system

Koivuranta *et al.*'s scoring system was derived from a dataset of 1107 surgical inpatients aged 4–86 years, who were prospectively included during a 3-month period in 1995 in one centre in Oulu, Finland [5]. A subset of these data was part of the combined data set from which Apfel *et al.*'s scoring system was derived. The original paper reported incidences of nausea and vomiting separately. Both outcomes were assessed for the intervals 0–2 h and 2–24 h after surgery, with nausea rated on an 11-point numeric scale (0–10). As the overall incidence of PONV 24 h after surgery was not reported, we estimated this using the incidence of nausea, which was 52% (Table 1, column 3). The formula of the original scoring system, developed using multivariable logistic regression analysis, was:

Table 1 Distribution of predictors of PONV and outcome (where presented) in the derivation [4, 5] and validation [18] datasets. Values are number (proportion), median (interquartile range) or median [range].

	Derivation datasets		Validation dataset
	Apfel *et al.* [4]* (*n* = 1040)	Koivuranta *et al.* [5] (*n* = 1107)	Visser *et al.* [18] (*n* = 1388)
Predictors			
Female gender	593 (57%)	731 (66%)	791 (57%)
Age; years	49 (36–61)	46 [4–86]	44 (32–56)
History of PONV	–	476 (43%)	347 (25%)
History of motion sickness	–	266 (24%)	264 (19%)
History of PONV or motion sickness	364 (35%)	–	527 (38%)
Non-smoker	759 (73%)	863 (78%)	847 (61%)
Duration of procedure; min	116 (73–176)	72 (40–105)	120 (80–173)
Duration of procedure > 60 min	–	–	1249 (90%)
Use of postoperative opioids	478 (46%)	786 (71%)	874 (63%)
Outcome			
Overall incidence of PONV	458 (44%)	576 (52%)	666 (48%)

*Values in this combined dataset were estimated by averaging the numbers presented for the two original datasets (Oulu and Wuerzburg).

$$\begin{aligned} Risk\ of\ PONV = 1/(1 + exp(-2.21 + 0.93 \times female\ gender \\ + 0.82 \times history\ of\ PONV \\ + 0.59 \times hisory\ of\ motion\ sickness \\ + 0.61 \times non\text{-}smoking \\ + 0.75 \times duration\ of\ surgery\ over\ 60\,min)). \end{aligned} \quad (2)$$

The intercept (−2.21) was obtained by means of written communication with the author. This baseline risk was about the same as in Apfel *et al.*'s scoring system.

Koivuranta *et al.*'s scoring system was also simplified to a five-item risk score, which was defined as the number of predictors present. If none, one, two, three, four or five of these predictors were present, the estimated predicted risk of nausea was 17%, 18%, 42%, 54%, 74% and 87%, respectively. Internal validation of the scoring system or risk score was not reported.

Validation dataset

The present (validation) dataset included surgical inpatients aged 18–80 years, who were enrolled in the 'AIDA' trial conducted at the Academic Medical Centre of the University of Amsterdam, the Netherlands, from April 1997 to January 1999 [18]. The objective of this trial was to assess the incidence of PONV after total intravenous anaesthesia with propofol *vs*. inhalation anaesthesia with isoflurane-nitrous oxide. The trial was approved by the institutional medical ethics committee and written informed consent was given by all patients. The design and primary results have been reported elsewhere [18].

The AIDA trial randomly assigned 1447 inpatients, for whom follow-up data at 24 h were complete in 1403 (97%). For the present analyses, another 15 patients were excluded from the sample because of missing values for one or more predictors included in Apfel *et al.*'s or Koivuranta *et al.*'s scoring system, leaving 1388 analysable patients. The distribution of the other predictors was similar among the 1388 patients with complete and the 59 patients with missing data. All types of surgery were included except cardiac surgery and intracranial procedures. The study population represented 1029 superficial (74%), 53 upper abdominal (4%), 81 lower abdominal (6%), 104 laparoscopic (8%), 12 strabismus (1%), 95 middle ear (7%) procedures, and 14 other procedures (1%). Other exclusion criteria were emergency surgery, (possible) pregnancy, ASA physical status 4, morbid obesity (weight > 120 kg), renal or liver disease precluding use of either anaesthetic technique, use of anti-emetic or pro-emetic medication in the 2 weeks before surgery, previous enrolment in the same study and insufficient command of the Dutch language. For all study subjects, we documented information on predictors included in Apfel *et al.*'s and Koivuranta *et al.*'s scoring systems, i.e. female gender, smoking behaviour, history of PONV and motion sickness, the type and duration of the scheduled surgery, the duration of anaesthesia and postoperative administration of opioid analgesics. All patients were interviewed (every 15 min in the recovery unit and hourly for 24 h after surgery in the ward) by a thoroughly trained research nurse who was blinded to the anaesthetic technique. The nurse recorded nausea, retching, vomiting and analgesic and anti-emetic medication. Patients were considered to have had PONV if at least one episode of postoperative nausea, retching or vomiting, or any combination of these, occurred in the first 24 h.

Data analysis

To validate the original scoring systems, the risk of PONV was calculated for each patient in the validation dataset using equations 1 and 2. To validate the two simplified risk scores, the number of predictors present was counted for each patient in the validation dataset. The predictive accuracy of the original scoring systems and simplified risk scores was quantified using measures of discrimination and calibration. Discrimination, the ability of a risk score or scoring system to distinguish between patients with and without PONV, was estimated with the area under the Receiver Operating Characteristic curve (ROC area). The ROC area may theoretically range from 0.5 (discrimination equivalent to that of chance) to 1.0 (perfect discrimination) [19]. Calibration refers to the agreement between the observed and the expected (predicted) outcome frequencies. For each original scoring system, calibration was studied by estimating a smoothed calibration line between the observed and predicted outcomes. Such a calibration line has an intercept and a slope which are ideally 1 and 0, respectively (perfect calibration). To estimate the slope of the calibration line in our validation dataset, we fitted a binary logistic regression model with the occurrence of PONV observed in the validation set as (binary) outcome and the linear predictor of the original scoring system as the only covariate [20]. This linear predictor was estimated for each patient in the validation dataset using eqns. 1 and 2 above. A patient's linear predictor is the sum (on the log scale) of the original regression coefficients multiplied by that patient's observed value of the corresponding predictor. The calibration of the original scoring systems was further tested with the Hosmer–Lemeshow test statistic [21]. For each simplified risk score, calibration was studied with a histogram, comparing the incidence of PONV in the validation and derivation datasets.

The clinical applicability of the risk scores was evaluated for several risk thresholds using conventional performance measures: positive predictive value, negative predictive value, sensitivity and specificity. For each threshold, the percentage of patients that would be treated with anti-emetic prophylaxis was also calculated.

Analyses were performed with SPSS 10.1 (SPSS Inc., Chicago, IL) and S-PLUS 2000 Professional Edition for WINDOWS.

Results

Table 1 shows the distribution of patients' characteristics, predictors and outcomes for the two derivation studies and the present validation study. The three datasets were largely comparable, except that the validation dataset included fewer patients with a history of PONV, fewer non-smokers, and an intermediate proportion of patients who received postoperative opioids.

Apfel *et al.*'s scoring system

The ROC area for Apfel *et al.*'s original scoring system was 0.62 in the validation dataset. We could not compare this with the ROC area of the scoring system found in the derivation dataset because this was not presented in the original publication. For the simplified risk score, the ROC area was 0.63 in the validation dataset, indicating that the simplified risk score and the original scoring system had a similar discriminative ability. The ROC area of the simplified score was larger in the derivation dataset than in the validation dataset (Table 2).

For Apfel *et al.*'s original scoring system (eqn. 1), the agreement between the predicted risk and the observed incidence of PONV is shown in Fig. 1a. The slope of the calibration line was 0.39, suggesting poor calibration. This was confirmed by a statistically significant Hosmer–Lemeshow test ($p < 0.001$). Application of the simplified risk score to the validation set yielded observed PONV incidences as shown in Fig. 2a. For patients with 0, 1 or 2 predictors, the observed incidences of PONV in the validation dataset were substantially higher than would have been expected based on the results from the derivation dataset, whereas the incidences of PONV were lower for patients in whom 3 or 4 predictors were present. Hence, for patients with a low risk score, the predicted risk of PONV was too low, whereas for patients with a high risk score, the predicted risk was too high.

Table 3a shows the number of patients with and without PONV in the validation dataset for each possible number of predictors according to Apfel *et al.*'s simplified risk score, as well as the characteristics of the risk score for several risk thresholds for administering anti-emetic prophylaxis. For example, using a threshold of three or more risk factors for defining a high risk of

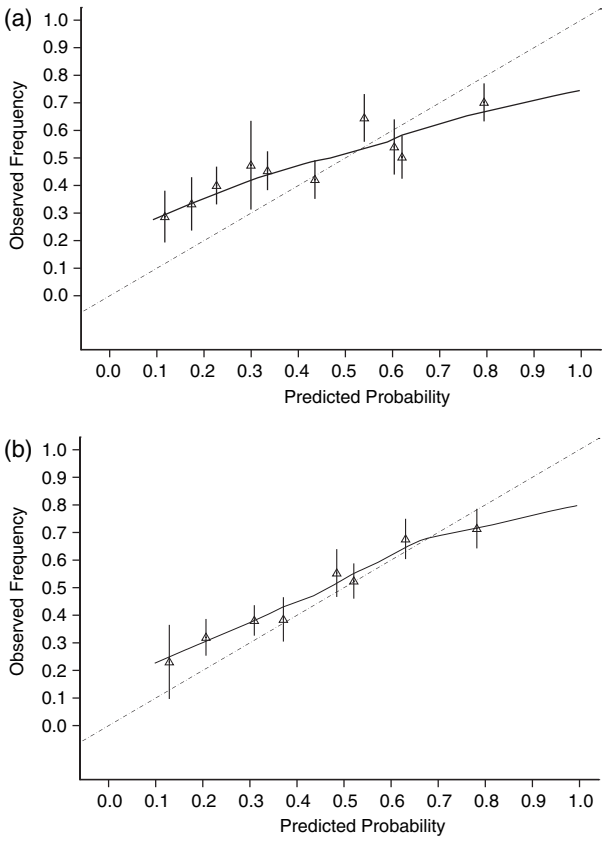

Figure 1 Calibration lines of the original scoring systems: a) Apfel *et al.*; eqn. 1 [4]; b) Koivuranta; eqn. 2 [5], in the validation dataset [18]. Triangles indicate the observed frequency of PONV per decile of predicted risk. The vertical lines represent the corresponding 95% CI. The solid line shows the relation between observed outcomes and predicted risks. Ideally, this line equals the dotted line that represents perfect calibration.

Table 2 Areas under the Receiver Operating Characteristic (ROC) curve for Apfel *et al.*'s [4] and Koivuranta *et al.*'s [5] risk scores as found in the respective derivation dataset and the validation dataset [18]. Ranges in brackets are 95% CI (where provided in the original paper).

	Derivation dataset	Validation dataset
Apfel *et al.*	0.75 (0.71–0.79)	0.63 (0.60–0.66)*
Koivuranta *et al.*		
Nausea	0.72	0.66 (0.63–0.69)
Vomiting	0.70	0.65 (0.62–0.68)
PONV	†	0.66 (0.63–0.69)

*p = 0.05.
†It may be assumed that the ROC area for PONV equals that for nausea, as the incidence of PONV commonly equals the incidence of nausea.

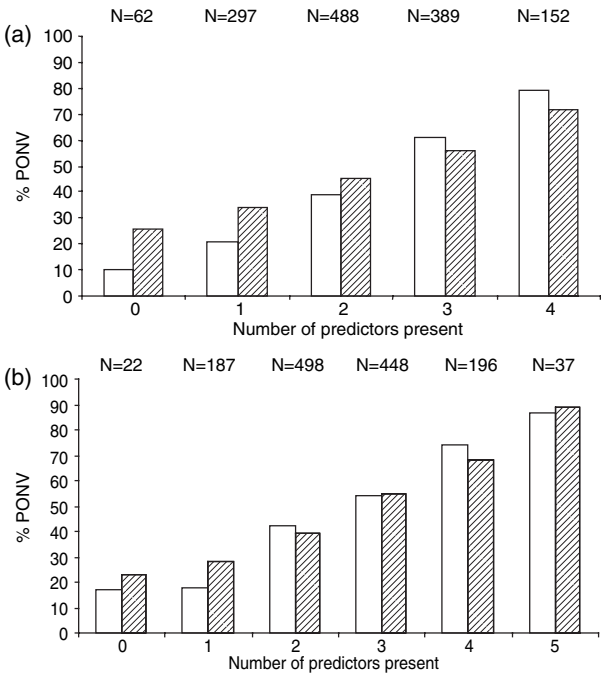

Figure 2 Percentage of patients with PONV per risk class as found in the derivation (clear) and validation (shaded) datasets: a) Apfel *et al.* [4]; b) Koivuranta *et al.* [5]. Figures at the top indicate the number of patients per risk class in the validation dataset [18].

PONV, 39% ((389 + 152)/1388) of all patients would receive anti-emetic prophylaxis. This would result in a positive predictive value ((217 + 110)/(389 + 152)) and a negative predictive value ((46 + 195 + 266)/(62 + 297 + 488)) of 60% each. Of all patients who developed PONV, 49% ((217 + 110)/667) would be correctly treated (sensitivity), but 30% ((172 + 42)/721) of all patients who did not develop PONV would receive unnecessary anti-emetic treatment (1-specificity).

Koivuranta *et al.*'s scoring system

For Koivuranta *et al.*'s original scoring system and for the simplified risk score, the ROC area for PONV was 0.66 in the validation dataset (Table 2). Unfortunately, the ROC areas could not be compared with those of the derivation dataset as the 95% CI and SE were not presented in the original paper.

For the original scoring system, the agreement between the predicted risk and the observed incidence of PONV is shown in Fig. 1b. The slope of the calibration line was 0.58 and the Hosmer–Lemeshow test was statistically significant ($p < 0.001$), also indicating poor calibration. Application of the simplified risk score to the validation dataset yielded the incidences of PONV given in Fig. 2b.

Table 3b shows the number of patients with and without PONV in the validation dataset for each possible number of predictors according to Koivuranta *et al.*'s simplified risk score. This table can be read similarly to Table 3a.

Discussion

In the present study, we tested the applicability of two scoring systems [4, 5] for predicting the risk of PONV within the first 24 h after surgery. The predictive accuracy of the scoring systems in the validation dataset was substantially lower than in the datasets that were used to develop the scoring systems. Both scoring systems provided risk predictions that were too extreme, i.e. the predicted low risks were too low and the predicted high risks too high. Also, the ability of the scoring systems to discriminate between the presence and absence of PONV was poorer than expected based on results from previous studies. Furthermore, none of the possible risk thresholds for anti-emetic prophylaxis delivered satisfactory results in terms of predictive values or sensitivity and specificity. Hence, in their original forms, the scoring systems do not guarantee accurate prediction of the risk of PONV in other patient populations.

Apfel *et al.*'s scoring system

In the validation dataset, the discrimination of the original scoring system and simplified risk score was poor. For the risk score, the ROC area was 0.75 in the derivation dataset but 0.63 in the validation dataset. In previous validation studies, ROC areas were 0.68 (same institution where the scoring system was derived) [16] and 0.71 (predominantly female patients from a cancer referral centre) [11]. Loss of discriminative ability is often observed when scoring systems are too optimistic, which means that in new patients, high risk estimates are too high and low estimates too low [15, 21, 22]. A decrease in discriminative ability can sometimes also be explained by a difference in the distribution of predictor values between validation and derivation datasets. In this case, there were indeed differences in the distribution of non-smokers (61% vs. 73%) and of postoperative opioid analgesia (46% vs. 67%) (Table 1). However, these differences were relatively small and it is unlikely that they explain such a large decrease in ROC area. Neither is it likely that the decrease in ROC area was caused by differences in the definition of outcomes, since these were the same in the derivation and validation datasets.

The calibration of the original scoring system and simplified risk score was poor in the validation dataset. The calibration line of the original scoring system presented with a flattened slope of 0.39 instead of 1.0

Table 3 Number of patients in the validation dataset [18] with and without PONV for each possible number of predictors according to risk classification with the simplified risk scores. Values are number or proportion. Results of possible prophylactic policies for several treatment thresholds are shown.

	No. of predictors present					
a) Apfel *et al.* [4]	**0**	**1**	**2**	**3**	**4**	**Total**
No. of patients with PONV	16	102	222	217	110	667
No. of patients without PONV	46	195	266	172	42	721
Total no. of patients	62	297	488	389	152	1388
Threshold for administering prophylaxis		≥1	≥2	≥3	≥4	
Proportion of patients receiving prophylaxis		96%	74%	39%	11%	
Positive predictive value		49%	53%	60%	72%	
Negative predictive value		74%	67%	60%	55%	
Sensitivity		98%	82%	49%	16%	
Specificity		6%	33%	70%	94%	

	No. of predictors present						
b) Koivuranta *et al.* [5]	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
No. of patients with PONV	5	53	194	248	134	33	667
No. of patients without PONV	17	134	304	200	62	4	721
Total no. of patients	22	187	498	448	196	37	1388
Threshold for administering prophylaxis		≥1	≥2	≥3	≥4	≥5	
Proportion of patients receiving prophylaxis		98%	85%	49%	17%	3%	
Positive predictive value		48%	52%	61%	72%	89%	
Negative predictive value		77%	72%	64%	57%	53%	
Sensitivity		99%	91%	62%	25%	5%	
Specificity		2%	21%	63%	91%	99%	

(perfect calibration). This indicates that in patients for whom the original scoring system predicted a low risk of PONV, the actual incidence of PONV was much higher than predicted, whereas in patients who were classified as having a high risk, the incidence was lower. The flattened slope confirms that the observed decrease in ROC area was caused by over-optimism of the original scoring system [15, 20–22]. This means that the strength of the association of certain predictors with the outcome was probably overestimated [15]. Indeed, we found that the odds ratio for female gender and postoperative opioids were significantly higher in the derivation dataset than in the validation dataset (Table 4). Moreover, postoperative opioid analgesia did not show a significant association with the outcome (OR 1.1) in the validation dataset, compared with an OR of 2.2 in the derivation study (Table 4). In the study by Koivuranta *et al.*, use of postoperative opioids was significantly associated with nausea (OR 1.7), but it was not included in the model as only the five strongest predictors were selected.

Table 4 Odds ratios for the predictors of Apfel *et al.*'s [4] and Koivuranta *et al.*'s [5] original scoring systems, as found in the derivation and validation [18] datasets. Ranges in brackets are 95% CI.

	Derivation dataset	**Validation dataset**
Apfel *et al.*		
Female gender	3.6 (2.5–5.2)	1.6 (1.3–2.0)
History of PONV or motion sickness	1.9 (1.4–2.7)	2.3 (1.8–2.9)
Non-smoker	2.1 (1.5–2.8)	1.7 (1.4–2.2)
Postoperative opioids	2.2 (1.7–2.9)	1.1 (0.8–1.3)
Koivuranta *et al.*		
Female gender	2.5*	1.7 (1.3–2.1)
History of motion sickness	1.8*	1.9 (1.4–2.6)
History of PONV	2.3*	2.2 (1.7–2.8)
Duration of procedure >60 min	2.1*	1.7 (1.4–2.2)
Non-smoker	1.8*	1.9 (1.3–2.8)

*The original paper presents regression coefficients but not the OR; these were calculated based on regression coefficients presented in SCORE 1 in the original publication.

In previous validation studies of Apfel *et al.*'s scoring system, the reported slopes of the calibration line were 0.82 [16] and 1.08 [11]. It should be noted that these values are difficult to interpret and to compare with our results. Rather than using a binary logistic model based on data of all individual patients, these slopes were estimated using weighted linear regression analysis, based on data that were clustered within risk groups. This method has two disadvantages. First, clustering of patients' data results in loss of information. Second, as PONV is a dichotom-

ous outcome, it is better to use a binary logistic regression model to fit the calibration line than a linear regression model, which assumes a continuous outcome [15, 20–22].

Koivuranta *et al.*'s scoring system

For the original scoring system, the ROC area in the derivation dataset was 0.72 for nausea and 0.71 for vomiting. For the simplified risk score, the ROC area in the derivation dataset was 0.72 for nausea and 0.70 for vomiting. When applied to the validation dataset, the ROC area was 0.66 (nausea), 0.65 (vomiting) and 0.66 (PONV) for both the original scoring system and the simplified risk score. These results are concordant with the results from a previous validation study of Koivuranta *et al.*'s scoring system, which also included a large number of patients (n = 1566) undergoing various types of surgery [16]. A third validation study among 1444 patients undergoing a variety of surgical procedures, however, reported ROC areas much closer to the results found in the derivation dataset (0.71 for PONV and 0.73 for vomiting) [17].

The calibration line of Koivuranta *et al.*'s original scoring system showed a flattened slope of 0.58 in our data, which again indicates that the predicted risks were too extreme. As with Apfel *et al.*'s scoring system, this may be due to an overestimation of predictor–outcome associations, such as for female gender (OR was 2.5 in the derivation data set vs. 1.7 in our dataset) (Table 4). Nevertheless, Koivuranta *et al.*'s scoring system was substantially less miscalibrated than Apfel *et al.*'s. In previous validation studies, the reported slopes of the calibration lines were 1.13 [16] and 0.99 [17]. However, the values of these slopes are again difficult to interpret because of the analytical method chosen. It would be interesting to see the slopes of the calibration lines in these two studies, when estimated with binary logistic regression analysis.

Clinical implications

It is statistically attractive to quantify the predictive accuracy of a clinical scoring system with measures of discriminative ability and calibration. For clinical practice, however, if only patients with a relatively high risk of PONV will be treated with prophylaxis, it is less interesting whether a predicted risk of for example 10% should actually be 20%. It is more important to define a clinically relevant threshold, for which the benefits of prophylaxis against PONV outweigh the costs and possible side-effects. Therefore, we presented the test characteristics of Apfel *et al.*'s and Koivuranta *et al.*'s risk scores (Table 3) for several risk thresholds for anti-emetic prophylaxis. In the present validation dataset, none of the thresholds delivered satisfying results. For example, when using Apfel *et al.*'s risk score with a threshold for administering prophylactic anti-emetics of three or more predictors, 39% of all patients will be treated with a sensitivity of 49% and a specificity of 70%. This means that, at most, half of the patients who develop PONV will be treated with prophylaxis. Alternatively, when lowering the threshold to two or more predictors, the sensitivity increases to 82%, but at the expense of the specificity, which decreases to 33%. Hence, 67% of patients who will not develop PONV will receive prophylaxis unnecessarily. Since administration of prophylactic anti-emetics may expose patients to side-effects [3, 23], this is not desirable.

Ideally, the predictive accuracy of a scoring system is robust across populations. In reality, however, a scoring system generally predicts less accurately in new patients than in the patients used to develop the scoring system. Therefore, clinicians considering the use of a scoring system have three options, which have recently been summarised in detail [22]. First, one can use an existing scoring system without modification, knowing that the identified predictors or the weights assigned to them (regression coefficients) may not be ideal for one's own patient population. Second, one can accept the predictors in the published scoring system, but adjust the scoring system to one's own patient population (recalibration). This is done by multiplying the intercept and/or regression coefficients of the scoring system with a correction factor (shrinkage factor) that is estimated from one's own population. Third, one can revise an existing scoring system, by re-estimating individual regression coefficients (model revision) and including new predictors (model extension).

For our population, both scoring systems were clearly miscalibrated. In such an instance, recalibration (method 2) is indicated. In cases where the predictor weights of the scoring system apply to the new patient population, it may be sufficient to adjust the intercept only. This is the case when the miscalibration of the scoring system affects all patients in a similar way (e.g. due to withholding of N_2O in one dataset but not another). It is also the case when the derived scoring system lacks a certain predictor that plays a role in the new population. However, when not only the intercept but also the predictor weights are different, the scoring system as a whole may be recalibrated to the new population. This is done by simultaneously adjusting both the intercept and the regression coefficients with an overall correction factor, as previously described [22, 24, 25]. Finally, when there are differences between the derivation and validation datasets that specifically affect patients with certain characteristics (e.g. more frequent administration of prophylactic anti-emetics to women) or when predictor definitions

differ (e.g. different definitions of smoking), overall recalibration may not be sufficient and re-estimation of individual regression coefficients may be required (method 3). If this does not improve the performance of the scoring system, the scoring system may be extended by including additional predictors [22, 24–26].

The development of robust scoring systems can perhaps be compared to cumulative meta-analyses for estimating the efficacy of a particular treatment. It starts with one study that develops a scoring system. This scoring system is then tested (validated) and, if needed, adjusted using newly available data to improve the accuracy of the scoring system across various clinical settings. In our institution, we have chosen to revise the existing PONV scoring systems by re-estimating individual coefficients and extending the scoring systems by including additional predictors. The results of this analysis have been published recently [27]. It should be noted, however, that adjusting a scoring system to local circumstances and populations (methods 2 and 3) requires individual patient data on both the predictors and the occurrence of PONV. If such empirical data are not available, Koivuranta *et al.*'s scoring system may provide the most accurate predictions of risk of PONV, as this scoring system appears to be more robust across populations than Apfel *et al.*'s, according to the present and previous validation studies. To prevent unexpected outcomes, one might first test the scoring system(s) in one's own population before using it in daily clinical practice to guide anti-emetic management.

Acknowledgements

J.E.vd.B. was supported by the Department of Perioperative Care and Emergency Medicine, University Medical Centre Utrecht, Utrecht, the Netherlands, and by Glaxo Smith Kline BV (GSK), Zeist, the Netherlands. We gratefully acknowledge the support by The Netherlands Organisation for Scientific Research (ZON-MW 917.46.360).

References

1 Kovac AL. Prevention and treatment of postoperative nausea and vomiting. *Drugs* 2000; **59**: 213–43.
2 Watcha MF, White PF. Postoperative nausea and vomiting. Its etiology, treatment, and prevention. *Anesthesiology* 1992; **77**: 162–84.
3 Sung Y-F. Risks and benefits of drugs used in the management of postoperative nausea and vomiting. *Drug Safety* 1996; **14**: 181–97.
4 Apfel CC, Läärä E, Koivuranta M, Greim C-A, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting. *Anesthesiology* 1999; **91**: 693–700.
5 Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997; **52**: 443–9.
6 Palazzo M, Evans R. Logistic regression analysis of fixed patient factors for postoperative sickness: a model for risk assessment. *British Journal of Anaesthesia* 1993; **70**: 135–40.
7 Gan TJ. Postoperative nausea and vomiting – can it be eliminated? *Journal of the American Medical Association* 2002; **287**: 1233–6.
8 Apfel CC, Korttila K, Abdalla M, *et al.* A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. *New England Journal of Medicine* 2004; **350**: 2441–51.
9 Pierre S, Corno G, Benais H, Apfel CC. A risk score-dependent antiemetic approach effectively reduces postoperative nausea and vomiting – a continuous quality improvement initiative. *Canadian Journal of Anesthesia* 2004; **51**: 320–5.
10 Biedler A, Wermelt J, Kunitz O, *et al.* A risk adapted approach reduces the overall institutional incidence of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2004; **51**: 13–19.
11 Pierre S, Benais H, Pouymayou J. Apfel's simplified score may favourably predict the risk of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2002; **49**: 237–42.
12 Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *Journal of the American Medical Association* 1997; **277**: 488–94.
13 McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII. How to use articles about clinical decision rules. *Journal of the American Medical Association* 2000; **284**: 79–84.
14 Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *New England Journal of Medicine* 1985; **313**: 793–9.
15 Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**: 361–87.
16 Apfel CC, Kranke P, Eberhart LHJ, Roos A, Roewer N. Comparison of predictive models for postoperative nausea and vomiting. *British Journal of Anaesthesia* 2002; **88**: 234–40.
17 Eberhart LHJ, Högel J, Seeling W, Staack AM, Geldner G, Georgieff M. Evaluation of three risk scores to predict postoperative nausea and vomiting. *Acta Anaesthesiologica Scandinavica* 2000; **44**: 480–8.
18 Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea and vomiting and economic analysis. *Anesthesiology* 2001; **95**: 616–26.
19 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29–36.

20 Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 1986; **5**: 421–33.
21 Hosmer D, Lemeshow S. *Applied Logistic Regression*. New York: John Wiley & Sons, 1989.
22 Steyerberg EW, Borsboom GJJM, Van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine* 2004; **23**: 2567–86.
23 Domino KB, Anderson EA, Polissar NL, Posner KL. Comparative efficacy and safety of ondansetron, droperidol, and metoclopramide for preventing postoperative nausea and vomiting: a meta-analysis. *Anesthesia and Analgesia* 1999; **88**: 1370–9.
24 Van Houwelingen HC. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine* 1995; **14**: 1999–2008.
25 Van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 2000; **19**: 3401–15.
26 Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation* 1999; **99**: 2098–104.
27 Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value in the prediction of postoperative nausea and vomiting? *Anesthesia and Analgesia* 2005: in press.

