British Journal of Anaesthesia **86 (6)**: 822–7 (2001)

What can be expected from risk scores for predicting postoperative nausea and vomiting?

C. C. Apfel*, P. Kranke, C.-A. Greim and N. Roewer

Department of Anaesthesiology, University of Würzburg, Josef-Schneider-Str. 2, D-97080 Würzburg, Germany

**Corresponding author*

Several risk scores have been developed to calculate the *probability* of postoperative nausea and vomiting (PONV). However, the power to discriminate which *individual* will suffer from PONV is still limited. Thus, we wondered how the number of predictors in a score affects the discriminating power and how the characteristics of a population—which is needed to measure the power of a score—may affect the results. For ethical reasons and to be independent from centre specific populations, we developed a computer model to simulate virtual populations. Four populations were created according to number, frequency, and odds ratio of predictors. Population I: parameters were derived from a previously published paper to verify whether calculated and reported values are in accordance. Population II: a gynaecological population was created to investigate the impact of the study setting. Populations III and IV: to meet ideal assumptions a model with up to seven predictors with an odds ratio of 2 and 3 was tested, respectively. The discriminating power of a risk score was measured by the area under a receiver operating characteristic curve (AUC) and an increase of more than 0.025 per predictor was considered to be clinically relevant. The AUC of population I was similar to those reported in clinical investigations (0.72). The study setting had a considerable impact on the discriminating power since the AUC decreased to 0.65 in a gynaecological setting. The AUC with the 'idealized' populations III and IV was at best in the range of 0.7–0.8. The inclusion of more than five predictors did not lead to a clinically relevant improvement. The currently available simplified risk scores (with four or five predictors) are useful both as a method to estimate individual risk of PONV and as a method for comparing groups of patients for antiemetic trials. They are also superior to single predictor models which are just using the patients' history of PONV or female gender alone. However, our analysis suggests that the power to discriminate which *individual* will suffer from PONV will remain imperfect, even when more predictors are considered.

Br J Anaesth 2001; **86**: 822–7

Keywords: vomiting, nausea; risk

Accepted for publication: January 2, 2001

The incidence of post-operative nausea and vomiting (PONV) is still about 25–30%.[1] Thus, innumerable studies have been carried out to investigate prophylactic antiemetic treatments.[2] Meta-analyses have shown that efficacy of prophylactic antiemetic strategies is limited and that the number-needed-to-treat (NNT) to prevent one patient from PONV is at best in the range of 5 when the basal event rate is high.[3–5] In addition, prophylactic compared with therapeutic antiemetic treatment does not improve patient satisfaction unless a high basal event rate is present.[6] Thus, prophylactic antiemetics appear only justified in patients at increased risk for PONV.[7 8]

In the past, high-risk patients were intuitively classified by reference to the past medical history of PONV or the type of surgery. Recently, risk scores have provided an objective risk assessment for PONV.[9–11] Several studies have shown that the risk assessments derived from such scores are robust enough to be also valid both in other hospitals and under different conditions.[11–14] However, the power to discriminate which *individual* will suffer from PONV remains limited (discriminating power). Thus, some centres are starting to develop more complex scores,[15] hoping to gain better results by the inclusion of more predictors. Unfortunately, the development and the validation of such scores requires a

large number of patients not receiving prophylactic antiemetics which may be ethically questionable if they are at high risk according to current risk scores. Thus, for ethical reasons and to be independent from centre specific populations, we created virtual populations to explore what can be expected from risk scores for predicting PONV by investigating how the number of risk factors and different study settings may affect their discriminating power.

Methods

The virtual populations were created using a computer model as described in more detailed in the appendix.

Population I was created to verify that this model of a virtual population (when based on parameters taken from a previous study on real patients) leads to similar results when compared with the previously reported area under the receiver operating characteristic curve (AUC) from the 'real world'.[11] This was achieved by considering the frequencies and odds ratio (OR) from the previous publication. Then, risk scores considering female gender alone, non-smoking, history of motion sickness or PONV (MSPONVhist), plus post-operative opioids and plus the consideration of the interaction of male gender and MSPONVhist were created (Tables 1 and 2). Its discriminating power was measured by the AUC.

To investigate how the AUC might be affected by a different study setting, a gynaecological setting (population II) was constructed so that female gender could no longer be used as a predictor. The population also included type of operation to quantify the benefit of an additional predictor as most scores were unable to identify the type of operation as an independent risk factor. For this, we were assuming that gynaecological laparoscopies are associated with an OR of 2.3, as previously described.[16]

Finally, for populations III and IV, risk scores considering 1–7 idealized predictors (frequency 50% and OR 2 and 3, respectively) were applied to approximate what can *at best* be expected from an *n*-predictor model (Tables 1 and 2). In summary, four populations with different characteristics were created (Table 1 and 2).

- Population I: Incidence of PONV, frequency, and OR of the predictors were taken from a previous study based on real patients.[11]
- Population II: Incidence, frequency, and OR were modified to represent a gynaecological setting.
- Population III: Incidence of PONV=50%, frequency of risk factors=50%, OR=2.0.
- Population IV: Incidence and frequency as simulation III, but OR=3.0

Characteristics to create the populations were derived from the literature. A systematic review of studies investigating predictors using logistic regression models revealed two publications analysing postoperative vomiting (POV),[13 17] and six analysing PONV.[9–11 15 16 18] One study from a pharmaceutical company was excluded,[18] because individual predictors (e.g. patients history of PONV) were not appropriately considered. A second study was not considered[15] for reasons described in a joint letter from the UK, Finland, and Germany.[19] The remaining studies with more than 1000 patients[10 11 13 16 17] reported ORs for female gender in the range of about 3 while the OR of the other predictors were in the range of 2 or less. Thus, the first assumption was that the OR of clinically relevant predictors for PONV are at best in the range of 2–3.

While some scores allow exact calculations of the probability using the coefficients derived from the logistic

Table 1 Frequency of PONV and predictors in the different populations

Population I (%)	Population II (%)	Population III (%)	Population IV (%)
PONV=34.5	PONV=60	PONV=50	PONV=50
Females=56.9	Females=100	1st Predictor=50	1st Predictor=50
Non-smoker=73.1	Non-smoker=80	2nd Predictor=50	2nd Predictor=50
MS-PONVhist=35.1	MS-PONVhist=50	3rd Predictor=50	3rd Predictor=50
Post-operative opioids=45.8	Post-operative opioids=50	4th Predictor=50	4th Predictor=50
Interaction=7.6	Laparoscopies=25	5th Predictor=50	5th Predictor=50
–		6th Predictor=50	6th Predictor=50
–		7th Predictor=50	7th Predictor=50

Table 2 OR of predictors in the different populations

Population I	Population II	Population III	Population IV
Females=3.57	Females=3.57	1st Predictor=2.0	1st Predictor=3.0
Non-smoker=2.05	Non-smoker=2.05	2nd Predictor=2.0	2nd Predictor=3.0
MS-PONVhist=1.92	MS-PONVhist=1.92	3rd Predictor=2.0	3rd Predictor=3.0
Post-operative opioids=2.18	Post-operative opioids=2.18	4th Predictor=2.0	4th Predictor=3.0
Interaction=2.14	Laparoscopies=2.3	5th Predictor=2.0	5th Predictor=3.0
–		6th Predictor=2.0	6th Predictor=3.0
–		7th Predictor=2.0	7th Predictor=3.0

models[9 13 17] two recent studies provided evidence that a simplification by just considering the number of binary predictors does not significantly impair the discriminating power and still provides an appropriate estimate of the individuals risk.[10 11] This led to the second assumption: that considering binary variables does not significantly alter the results.

Palazzo and Evans found an interaction between gender and history of PONV.[9] This was also found in a cross-validation between two centres but detailed analysis revealed this effect to be negligible.[11] Because the other studies[10 13] did not find any other relevant interactions, the third assumption is that an interaction (covariation) between risk factors for PONV is negligible.

The discriminating power of a score was measured by the area under the receiver operating characteristic (ROC)-curve (AUC) as previously described.[10 11 13–15 17] When the average risk of PONV is 25% and the score results in a probability of 25% for almost every patient, then it would be impossible to discriminate *individuals* in that population who will suffer from those who will not. If, in contrast, 75% of the patients are predicted to have a relatively low risk while the remaining 25% will be predicted with a relatively high risk, then the score may predict PONV correctly in most of the *individuals*, for example, with acceptable discriminating power. Details on the calculation and interpretation of the discriminating power—which considers the relationship between sensitivity and specificity in the ROC-curve—are described elsewhere.[20 21]

For practical reasons, an increase in the AUC of 0.025 per additional risk factor is considered as clinically relevant. This is roughly associated with a 5% higher sensitivity at a specificity of 50%. As this modelling is based on virtual populations, confidence intervals or calculations of *P*-values are not appropriate.

Results

In population I, a single predictor model resulted in AUCs of about 0.60. Female gender (AUC=0.64) had a greater impact than non-smoking (AUC=0.57), history of motion sickness or PONV (0.58) or post-operative opioids (0.60) because of the high OR of 3.6 and the frequency or approximately to 50% (Fig. 1). The four-predictor model resulted in a clinically relevant higher AUC of 0.72 (Figs 1 and 2) which is similar to the average AUC of the cross-validated scores from a previous publication.[11] When an additional predictor (e.g. interaction between male gender and MSPONVhist) is considered (Fig. 2, upper graph), no clinically relevant improvement is noticeable (ΔAUC<0.01).

When such a score is applied to the gynaecological setting (population II) female gender did not contribute to the discriminating power of the score (AUC=0.5). Consideration of the other three predictors leads to an AUC of 0.65 (Fig. 2, lower graph). However, further predictors (e.g. laparoscopy with an OR of 2.3) did not improve discriminating power significantly (ΔAUC<0.025, Fig. 2).

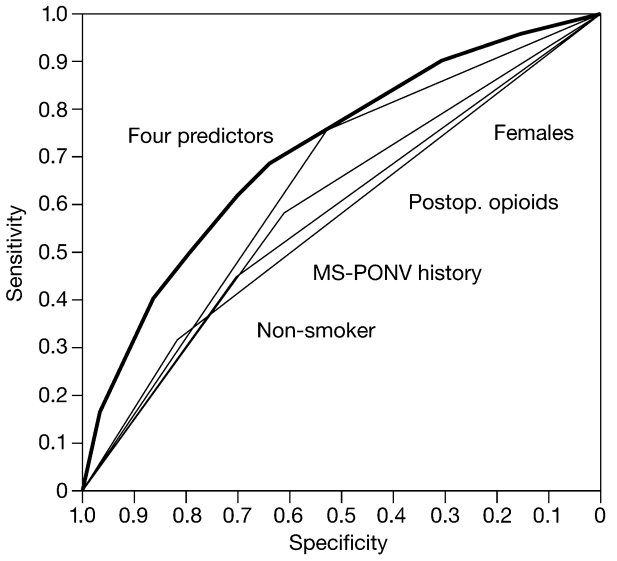

Fig 1 ROC-curve for the single predictor models and for the four-predictor model. The single predictor model considered either female gender, non-smoking, history of motion sickness or PONV and post-operative opioids while the four-predictor model considered all predictors simultaneously. Note, that the four-predictor model had a clinically relevant larger AUC than any single predictor model.

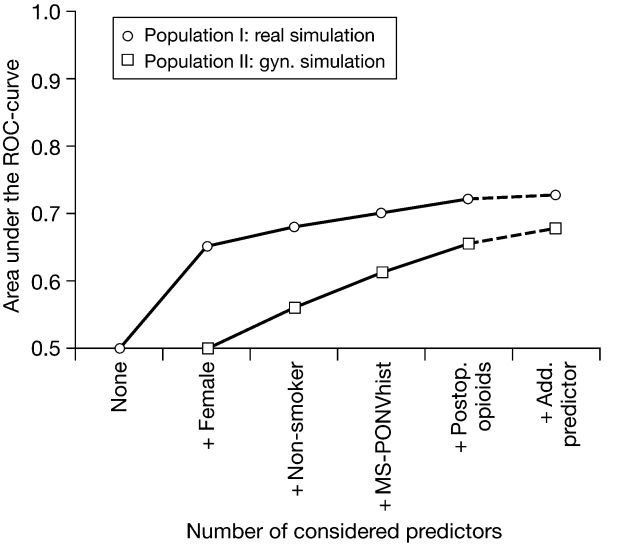

Fig 2 (Upper curve) AUC for population I, for example, when all parameters were taken from a previous study.[11] The interval indicated by the vertical lines displays the range of the published AUC when four or five predictors were applied to its original data set. (Lower curve) AUC for population II, for example, when this score is applied to a gynaecological setting. Note, that considering the additional predictor such as 'laparoscopy' does not lead to a clinically relevant improvement. The dotted lines indicate that the increase of the AUC per additional predictor is less than 0.025. MS-PONV hist=history of motion sickness or PONV. ROC-curve=receiver operating characteristic-curve.

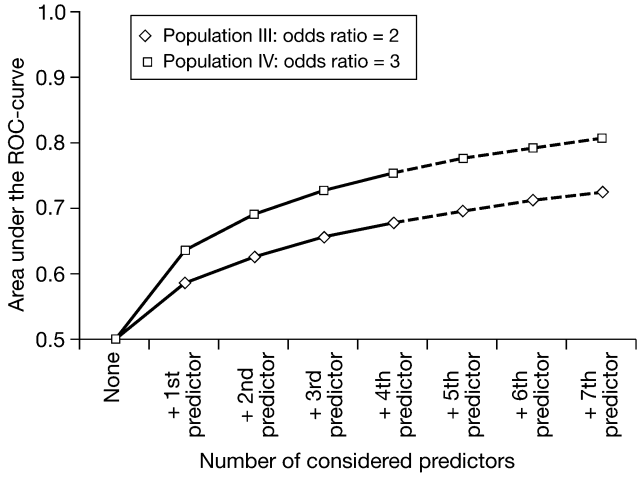

Fig 3 (Lower curve) AUC for population III, for example, all predictors were assumed to have an OR of 2. (Upper curve) AUC for population IV, for example, all predictors were assumed to have an odds ratio of 3. Note, that a four-predictor model is much better than a single factor prediction. However, from four or five factors onwards, the clinical relevance becomes questionable as the increase of the area under the curve per additional predictor becomes less than 0.025, as indicated by the dotted lines. ROC-curve=receiver operating characteristic-curve.

The area of the ROC-curves for populations III and IV, when 0–7 risk factors are considered, are depicted in Figure 3. Again, the improvement in discriminating power is most obvious when the first predictor is introduced and decreases with every additional factor. The data demonstrate that four predictors are clearly superior to a single predictor model (ΔAUC ~0.12) but the increase of the AUC with any further risk factor becomes less than 0.025.

Discussion

This virtual population model appeared to allow centre independent investigations to quantify the relative impact of risk factors on the discriminating power of scores for predicting PONV. It could be shown that the prediction with several risk factors is superior to a prediction with a single factor (e.g. female gender or the history of PONV) and resulted in an AUC of ~0.72. This corresponds to a sensitivity and specificity of about 75 and 50%, respectively. Therefore, we recommend the use of one of the simplified multifactorial models[10 11] for an individual risk assessment in daily practice. Further, these scores should be used for group comparisons in antiemetic trials as the impact of several non-significant risk factors may well add up to a clinically significant different baseline risk.[22] Although, their predictive ability remain imperfect, the suggested models (considering real and ideal parameters from current data of the literature) give little reason to assume that future models will lead to significantly better predictions unless other risk factors with much higher ORs are discovered and confirmed by several centres.

Population I, created from parameters of a previously published population, resulted in a discriminating power, as expressed by the area under the receiver operating characteristic curve (AUC), of 0.72 when four predictors were considered. This is in the range of 0.683–0.746 when a score was applied to its original centre.[11] As the discriminating power of the risk scores is comparable when applied to the virtual and the real population, we conclude that this model is a good representative of reality. Although the characteristics to create the virtual population were identical to the real population, other underlying characteristics were of course not considered. If, for example, a relatively strong, up to now, undiscovered predictor was unevenly distributed in the real but not in the virtual population, the discriminating power may have resulted in lower AUC in the real compared with the virtual population. Thus, it was reassuring that the AUC applied to the virtual population was in the expected range. Further, it could be shown that female gender as the sole predictor already has an AUC of 0.64, but this can be improved to 0.72, when three additional predictors are considered. However, any further predictor, such as an interaction between male gender and MSPONVhist, did not lead to a significant improvement (ΔAUC <0.025). For clinical purposes a score needs to be easy applicable and as addition of further risk factors may complicate calculations for little improvement in risk assessment, the benefit of introducing a further predictor needs to be justified.

To investigate the potential impact of a different study setting on the discriminating power of a score, we constructed the most extreme deviation from the original, which is a gynaecological setting. This would eliminate the benefit of the strongest predictor (population II). The discriminating power with the remaining three predictors was 0.65. This lower discrimination was to be expected. It is noteworthy that in an investigation of a very homogeneous population, for example, all patients are female non-smokers with MSPONVhist undergoing procedures, which most likely will require post-operative opioids, a score based on those predictors will probably calculate for every patient a risk of 79%.[11] As all patients will have the same risk, the score will not be able to discriminate between expected vomiters or non-vomiters, for example, AUC will be 0.5. This does not necessarily mean that the score is useless in that setting as it may still indicate that all these patients have a very high risk for PONV, which could justify the use of prophylactic antiemetics. This example demonstrates that the discriminating power of a scoring system may be affected by the investigated population and that the calibration curve is another important descriptor which can not be analysed with this model. In this respect, some validation studies may be needed to provide acceptable calibration curves in other centres. Interestingly, the risk models which have been validated from other centres[10 11] were independent of the type of operation as its relative impact was negligible. To the best of our knowledge, there

seem to be only two valid studies in which the type of operation led to statistically significant OR in multivariate analyses.[13 16] However, their results are controversial and a score considering the operation did not lead to a better prediction compared with the previously described operation-independent model.[13] Although there is little evidence supporting the widespread notion that the type of surgery is a strong risk factor, we tested this hypothesis. Given that 25% of gynaecological patients will undergo laparoscopic surgery with an OR of 2.3,[16] this improvement of about 0.02 led to an AUC of 0.67 which does not appear to be clinically relevant.

The best results can be expected if the frequency of PONV is 50% and if the frequencies of the predictors are also 50%. As pointed out, the overall OR can at best be assumed to be in the range of 2–3.[10 11 16 17] These assumptions were considered in the virtual populations III and IV and resulted in AUCs in the range of 0.72 up to 0.8 which support the previously established models reporting AUCs between 0.71[10] and 0.78[17] in real populations. Again, further inclusion of more than four or five risk factors does not lead to a clinically relevant increase in the discriminating power in the population models as found by Koivuranta and coworkers who reported an AUC of 0.721 in an eight factors and 0.710 in a five factor model.

Creating a virtual population led to practically identical results to those reported from real populations. The discriminating power of risk scores for predicting PONV can at best be expected to be in the range of 0.7–0.8 which means that the discrimination of *individuals* who will suffer from those who will not is still imperfect and will not significantly improve with further predictors. Unless there is consistent evidence that other predictors with a much stronger impact do exist, it is unlikely that future risk scores will provide a significantly better prediction for PONV. Thus, for the time being, it may be ethically questionable to develop new risk scores based on a large number of patients known to be at high risk who would be deprived of an effective prophylactic antiemetic strategy.

Appendix

Creation of the virtual population

For each predictor the coefficients were calculated by taking the logarithm of the OR. Up to seven predictors were considered in a model so that 2^7=128 combinations are possible.

The *conditioned frequency* of a combination was calculated from the product of the single frequencies of the predictors. The value was multiplied by 100 and the integer was taken. This results in the *number of patients in that combination* which is a proportional representation of that combination in the population.

The *conditioned probability of PONV* ($P_{(PONV)}$) for that combination was calculated by the sum of the coefficients according to the presence of the predictors and submitted to a logit transformation $P_{(PONV)}=1/(1+e^{-(sum\ of\ coefficients)})$.

The number of patients with PONV was derived from the number of patients in that combination multiplied by the conditioned probability.

All *patients in that combination* were added (=total number of patients=population). All *patients with PONV* were added and divided by the total number of patients (=incidence of PONV). A constant was fitted into the calculation of the conditioned probability of PONV until the aimed incidence of PONV was reached (regression process).

References

1 Kovac AL. Prevention and treatment of postoperative nausea and vomiting. *Drugs* 2000; **59**: 213–43

2 Rowbotham DJ. Current management of postoperative nausea and vomiting. *Br J Anaesth* 1992; **69**: 46S–59S

3 Tramèr MR, Moore RA, Reynolds DJ, *et al.* A quantitative systematic review of ondansetron in treatment of established postoperative nausea and vomiting. *BMJ* 1997; **314**: 1088–92

4 Sneyd JR, Carr A, Byrom WD, *et al.* A meta-analysis of nausea and vomiting following maintenance of anaesthesia with propofol or inhalational agents. *Eur J Anaesthesiol* 1998; **15**: 433–45

5 Figueredo ED, Canosa LG. Prophylactic ondansetron for postoperstive emesis. Meta-analysis of its effectiveness in patients with previous history of postoperative nausea and vomiting. *Acta Anaesthesiol Scand* 1999; **43**: 637–44

6 Scuderi PE, James RL, Harris L, *et al.* Antiemetic prophylaxis does not improve outcomes after outpatient surgery when compared to symptomatic treatment. *Anesthesiology* 1999; **90**: 360–71

7 White PF, Watcha MF. Postoperative nausea and vomiting: prophylaxis versus treatment [editorial]. *Anesth Analg* 1999; **89**: 1337–9

8 Watcha MF. The cost-effective management of postoperative nausea and vomiting [editorial]. *Anesthesiology* 2000; **92**: 931–3

9 Palazzo M, Evans R. Logistic regression analysis of fixed patient factors for postoperative sickness: a model for risk assessment. *Br J Anaesth* 1993; **70**: 135–40

10 Koivuranta M, Läärä E, Snare L, *et al.* A survey of postoperative nausea and vomiting. *Anaesthesia* 1997; **52**: 443–9

11 Apfel CC, Läärä E, Koivuranta M, *et al.* A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology* 1999; **91**: 693–700

12 Toner CC, Broomhead CJ, Littlejohn IH, *et al.* Prediction of postoperative nausea and vomiting using a logistic regression model. *Br J Anaesth* 1996; **76**: 347–51

13 Apfel CC, Greim CA, Haubitz I, *et al.* The discriminating power of a risk score for postoperative vomiting in adults undergoing various types of surgery. *Acta Anaesthesiol Scand* 1998; **42**: 502–9

14 Eberhart LH, Hogel J, Seeling W, *et al.* Evaluation of three risk scores to predict postoperative nausea and vomiting. *Acta Anaesthesiol Scand* 2000; **44**: 480–8

15 Sinclair D, Chung F, Mezei G. Can postoperative nausea and vomiting be predicted? *Anesthesiology* 1999; **91**: 109–18

16 Cohen MM, Duncan PG, DeBoer DP, *et al.* The postoperative interview: assessing risk factors for nausea and vomiting. *Anesth Analg* 1994; **78**: 7–16

17 Apfel CC, Greim CA, Haubitz I, *et al.* A risk score to predict the

probability of postoperative vomiting in adults. *Acta Anaesthesiol Scand* 1998; **42**: 495–501

18 Haigh CG, Kaplan LA, Durham JM, *et al*. Nausea and vomiting after gynaecological surgery: a meta-analysis of factors affecting their incidence. *Br J Anaesth* 1993; **71**: 517–22

19 Apfel CC, Palazzo M, Koivuranta M, *et al*. Models for predicting PONV are well established. Data acquisition and analysis bias of a new model may not add to current knowledge. *Anesthesiology* 2000; **92**: 1489–92

20 Hanley JA, McNeil BJ. The meaning and use of the area under a ROC curve. *Radiology* 1982; **143**: 29–36

21 Hanley JA, McNeil BJ. A method of comparing the area under receiver operating chacteristic curves derived from the same cases. *Radiology* 1983; **148**: 839–43

22 Matson A, Palazzo M. Postoperative nausea and vomiting. In: Adams AP and Cashman JN (eds). *Recent Advances in Anaesthesia and Analgesia*, Vol. 19. London: Churchill-Livingstone, 1995; 107–126