Quantifying prognosis with risk predictions
Nathan L. Pace, Leopold H.J. Eberhart and Peter R. Kranke

Introduction
Anaesthesiologists have been pioneers in the development and use of risk scores and risk prediction. Apgar1 was an experienced anaesthesiologist and an astute clinician. On the basis of the author’s careful observations of thousands of newborns, the author proposed in 1953 a ‘new method of evaluation’ that became the Apgar score (five signs summing up to a score of 10). Infants had mortality proportional to their Apgar score: 14% – Apgar score 0–2, 1.1% – Apgar score 3–7, and 0.1% – Apgar score 8–10. Fifty years later, the Apgar score is just as relevant in predicting neonatal mortality: 24% – Apgar score 0–3, 9.9% – Apgar score 4–6, and 0.02% – Apgar score 7–10.2 Even earlier, in 1941, the American Society of Anaesthesiologists promulgated a physical status score (the ASA score) of patients prior to anaesthesia and surgery.3 Although specifically not claiming to predict operative risk, the ASA score has in fact been demonstrated to be a risk score with a probabilistic interpretation for mortality and morbidity.4

Medical diagnosis, the identification of the nature and cause of a condition or event, reflects the current vitality of the patient – their present condition. A numerical Apgar score is a diagnosis of neonatal vigour; from low to high, it triggers a spectrum of resuscitation from aggressive efforts to simple observation. The Apgar score is also a risk score for prognosis. Prognosis is to know before or to give a forecast of the probable course and outcome of a disease, a procedure, a drug and so on. The baseline risk of an outcome for a population can come from several sources including: large randomised controlled trials; observational studies; and retrospective analysis of hospital or administrative databases. Multiple studies of each type may be assembled into a meta-analysis. The first very large study of perioperative mortality was a review of medical records at 10 US university hospitals. Records of 600 000 patients during 5 years (1948–1952) uncovered 8 000 deaths for an inhospital rate of 1.3% or one death for each 75 surgeries; of these, expert opinion classified only 224 deaths as anaesthetic related.5 Fifty years later, the US Medicare Provider Analysis and Review database (MEDPAR) with 35 million surgical patients from 2001 to 2006 demonstrated an in-hospital mortality of 3.1% or one death for each 32 surgeries.6 It is unlikely that this higher mortality reflects a deterioration of care, but is the result of an aging population, a different mix of surgical procedures, an increased prevalence of concomitant diseases, the redirection of healthy patients to ambulatory surgery and so on. Clearly, some risk probabilities only reflect earlier or different times.

A complete prognosis includes the expected time course and outcomes of the disease or surgery – the extreme outcomes being life or death. Risk is the potential that a disease or an action will lead to an undesirable outcome. Risk is always a probability issue – probability ($P_i$) being between 0 and 1. Probability reflects the continuum...
between absolute certainty \((P = 1)\) and certified impossibility \((P = 0)\). The future of individual patients cannot be foretold, but the probable outcome for a group of similar patients might be foretold. Using the MEDPAR data, an older patient could be informed preoperatively that their \(P\), for death is 0.031 (by assigning each patient the baseline risk for US Medicare patients). This is an unsatisfying strategy. The desire to define categories of patients with similar risk has prompted the exploration of risk factors and probabilistic risk predictions. Reports on the quantification of risk have now become commonplace. The methods usually require statistical tools that are not familiar to most anaesthesiologists. Additionally, these statistical tools can be unintentionally misused, the results misunderstood and biased descriptions of risk reported. We will briefly describe methods for identifying risk factors and risk scores, including the use of statistical regression techniques and a framework for assessing the performance of prediction models. This is not a systematic explanation or description of any particular use of risk factors, and to assist the reader illustrative examples will be cited.

### Risk factors

#### Selection

Observational data has become a fertile source of material for epidemiological reports; these patient records have been collected on diseases, treatments, events, outcomes and so on by hospital and government information systems. There are methodological shortcomings in primary studies of prognosis based on observational sources because of problems with the data. These potential biases have been previously described and include failure to clearly define and describe the source population and failure to adequately measure the putative prognostic factors.\(^7\) Apgar used clinical judgment to derive the Apgar score. Some identification of risk factors is still achieved by consensus panels of expert physicians, such as the Berlin Questionnaire for the identification of obstructive sleep apnoea.\(^8\) Simple classification of outcomes into contingency tables was used in the National Halothane Study to relate physical status and anaesthetic technique to the standardised death rate.\(^9\)

Current practices for identifying risk factors are generally numerical methods (Table 1);\(^15-19\) the most commonly used is logistic regression. About 50 years ago an algorithm was proposed for an automatic procedure to select a statistical model (i.e. to choose risk factors) in which there are a large number of potential risk factors and no underlying theory on which to base the model selection, this is the ‘stepwise algorithm’.\(^10\) The statistical notation is given by the multivariable, linear logistic model equation:  
\[
\ln \left( \frac{P_i}{1-P_i} \right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}
\]

The \(x_{j,i}\) represent the covariates that might be risk factors; these observed covariates are the independent (explanatory) variables speculated to be risk related. They may be dichotomous, for example, gender, with coding \(x_{j,i}\), of 0 or 1. The Greek letter \(\beta\) is the regression (\(\beta\))-coefficients. This regression equation is estimated multiple times; at each step the calculated values of some or all of the \(\beta\)-coefficients may change. This model \(\{\text{i.e. the covariates chosen and the regression coefficients estimated}\}\) is selected to maximise the goodness of fit of covariate values to the presence or absence of the event. An automatic variable selection process adds (forward selection) or removes (backward elimination) covariates from the model by a statistical measure of the model fit. When succeeding steps no longer include or exclude covariates, the stepwise process stops and reports the identity of the remaining covariates and the \(\beta\)-coefficients in the model. These remaining covariates are commonly called the ‘independent predictors’ or the ‘independent risk factors’. The ‘independence’ of these risk factors is frequently misunderstood: this is a purely statistical concept – if the beta coefficient divided by its standard error is more than 1.96, it is significant at \(P\) value of 0.05; the statement of statistical significance for each risk factor is only valid within the context of the total set of final risk factors and within that particular statistical model; the association of risk factors to outcome does not imply causality; and covariates that fail inclusion in the stepwise model may nevertheless be causal factors.\(^20\)

Multiple problems with the statistical methodology of the usual application of stepwise regression have been identified. Prognosis studies may include a large number of covariates and as each covariate is added to the stepwise regression, the number of possible models increases by a factor of two. With 50 covariates there are \(2^{50}\) (approximately \(10^{15}\)) or so possible candidate models. Some risk factors may be synergistic, for example smoking and hypertension. To include the possibility of synergy and antagonism between risk factors, first order interactions

### Table 1 Methods for selection of risk factors

<table>
<thead>
<tr>
<th>Method type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical expertise or experience</td>
<td>Apgar score(^1)</td>
</tr>
<tr>
<td>Classification tables</td>
<td>National Halothane Study(^2)</td>
</tr>
<tr>
<td>Multivariable logistic regression (all covariates)(^10,11)</td>
<td>VA day-of-surgery deaths(^3,12)</td>
</tr>
<tr>
<td>Propensity analysis(^14)</td>
<td>Chronic versus acute β-blockade after non-cardiac surgery(^4)</td>
</tr>
<tr>
<td>Machine learning (artificial neural network)(^15)</td>
<td>PONV prediction(^16)</td>
</tr>
<tr>
<td>Bayesian logistic regression(^15,16)</td>
<td>Morbidity after coronary artery surgery(^15)</td>
</tr>
</tbody>
</table>

\(^{PONV}, \text{postoperative nausea and vomiting; VA, Veterans Administration.}\)

---

**European Journal of Anaesthesiology** 2012, Vol 29 No 1

Copyright © European Society of Anaesthesiology. Unauthorized reproduction of this article is prohibited.
Interpreting this very large OR is made difficult in a systematic review et al. values less than 0.05 can be created. Brain More fundamentally, stepwise regression will create sets of risk factors composed of both the true and the false; they cannot be distinguished from each other. If large sets of variables generated by a random number process are subjected to stepwise regression, the appearance of association will usually result. From noise, independent predictors with publishable $P$ values less than 0.05 can be created.

Another weakness is the inconsistent analysis and reporting of potential risk factors. In 2005, Brotman et al. reported over 100 risk factors for cardiovascular outcomes in 117 published studies: ‘any claim that variable X is an independent risk factor for a given cardiovascular outcome (except within the context of a particular study) ignores the likelihood of residual confounding i.e. that valuable predictors also associated with X have been excluded, poorly measured, or incorrectly modelled.’ As another example, Ip et al. in a systematic review found 48 studies (23,037 patients) reporting risk factors for postoperative pain and analgesic consumption. Pre-existing pain, anxiety, age and type of surgery were the four most important risk factors. Yet of the 32 (of a total of 48) studies reporting on postoperative pain, the majority did not report results for these risk factors; only eight, 15, 12 and six studies reported on pre-existing pain, anxiety, age and surgery type as risk factors. Were these risk factors not analysed in most primary studies and, thus, could not be reported? Were these risk factors actually analysed in most primary studies and, thus, not reported because no association with pain intensity was found? The answers to these questions are not known.

Advances in statistical theory and software have provided alternatives to stepwise regression for choosing risk factors. These include propensity analysis, Bayesian approaches to logistic regression and the various approaches of machine learning, such as decision tree learning and artificial neural networks; all these have been used in topics of anaesthetic interest (Table 1). The incorporation of prior knowledge about the outcome in choosing risk factors is a particularly interesting aspect of Bayesian methods and in simulations has better statistical properties. The poor generalisability of risk factors from stepwise regression when applied to new patients is in part due to insufficient numbers of patients in the original data; the result is called an over-optimistic model. Attempts to correct for this overfitting by applying penalised likelihood, shrinkage factors, bootstrapping and so on now appear in anaesthesia journals. Regardless of the method for derivation, there must be validation and replication of risk factors.

Biomarkers
A biomarker is a naturally occurring substance used as an indicator of a biological state. Included within biomarkers are plasma enzymes, antibodies, chromosomal rearrangements, gene expression and so on. Biomarkers are chosen because their presence can be identified in a particular pathological or physiological process, disease and so on. Frequently a biomarker indicates a change in the expression, concentration or state of a protein that correlates with the risk or progression of disease. Enormous effort has been spent on searching for biomarkers of cancer. Many associations of a single or multiple proteins with different cancers have been reported. However, efforts to use these biomarkers as risk factors have often floundered because of the difference between statistical significance and prognostic discrimination. Brain natriuretic peptide (BNP) and troponins have been of special interest for anaesthesia and surgical risk. As an example, BNP is a risk factor for major adverse cardiovascular events (cardiac death or nonfatal myocardial infarction) after non-cardiac surgery. BNP is a circulating hormone synthesised by cardiomyocytes in response to increased ventricular wall stress or ischaemia and has natriuretic and vasodilator properties. In a systematic review of 15 studies (about 5,000 patients having non-cardiac surgery) using study level aggregate data, elevated BNP obtained preoperatively had a highly elevated odds ratio (OR) of 19.77 [95% confidence interval (CI) 13.18–29.65] for short-term major adverse cardiovascular events. Interpreting this very large OR is made difficult by the lack of a common definition for the threshold value to separate BNP into normal and abnormal ranges, a frequent problem in systematic reviews of prognosis. One generally accepted range of normal plasma BNP is up to 100 pg ml$^{-1}$; in the 15 studies, the threshold separating normal BNP from abnormal BNP ranged from as low as 35 pg ml$^{-1}$ to as high as 255 pg ml$^{-1}$. When a large sample of individual patient data can be used a better definition of the relationship between biomarker and outcome becomes possible. In a prospective observational study (>4,000 patients) of cardiovascular disease, higher serum cardiac troponin T concentrations were a predictor of increased risk of heart failure and cardiovascular mortality in older adults. In a meta-analysis using individual patient data from seven studies (about 19,000 patients), the creatine kinase (CK-MB) ratio (the ratio between the peak CK-MB and the
upper limit of normal of CK-MB) was calculated for each patient following coronary artery bypass grafting. Thirty-day mortality increased with increasing CK-MB; the relative risk was over four-fold higher for patients having a CK-MB ratio of 10–20 compared with those with a CK-MB ratio less than 1.35

**Multivariable probabilistic risk predictions**

The relationship between outcome and a risk factor may be examined individually as described above. As risk factors may interact, the effects of multiple risk factors on outcome are analysed together in order to avoid and minimise spurious correlations. This is performed by multivariable regression, a statistical tool for determining simultaneously the unique contributions of various factors to a single event or outcome. The performance of these statistical models should be validated for their overall performance, discrimination and calibration.

**Estimation**

For estimation(4,7),(990,995) the strength of association between an outcome and a dichotomous risk factor is usually specified by the OR, not by the probability value of a statistical test. If the risk \( P \) is 0.1, then for every 100 patients the event occurs in 10 patients. Odds is a concept from gambling, but is used also in medicine. If there are 10 patients with an event and 90 without an event, the odds are one to nine (written as the ratio 1 : 9 or as the decimal fraction 0.111...). Risk and odds are mathematically convertible: risk = odds/(1 + odds); odds = risk/(1 - risk). In an unadjusted OR only two variables, the binary outcome and the binary predictor, are used for the calculation (Table 2). The OR is the change in the odds with the presence of risk factor. An OR can also be calculated for a continuous risk factor such as age, and the OR represents the increase in odds of outcome for a one-unit change in the risk factor. The calculation of the 95% CI is also straightforward. At an OR of 1 (the identity value), there is no association of risk factor to outcome. If the OR value is greater or less than 1, there is still no association if the upper and lower bounds of the 95% CI lie on opposite sides of the line of identity. By contrast, if the bounds of a 95% CI are firmly above or below the identity value, the OR has achieved statistical significance. There is a distinction for continuous risk factors having an OR sufficient to demonstrate association versus an OR magnitude necessary for useful prognostic discrimination; OR must be much larger for clinical usefulness.32 An adjusted OR is a statistical method for isolating each risk factor’s independent effect within the total set of risk factors from a multivariable model. The statistical methods for adjusted ORs are considerably more complicated and require iterative solutions to multivariable logistic regression equations.35

A probability prediction rule assigns a probability \( P_i \) (called a prediction, a forecast or a prognosis) for the occurrence of a specified event to a patient; these predictions are mainly descriptive, not mechanistic or explanatory, of the associations between risk factors and outcome. Using essentially the same mathematical methods (stepwise logistic regression) as for identifying risk factors, a statistical model of risk factors is chosen:

\[
\log \left( \frac{P_i}{1 - P_i} \right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij},
\]

The adjusted OR for each risk factor is obtained by exponentiating the \( \beta \)-coefficient: \( e^\beta \) (Table 2). The ‘z’ is the risk score for the model; it is also called a risk index score or a risk index. Inherent in the model is the calculation of \( P_i \) for a new patient with specific risk factor values \( (x_{ij}) \). This is calculated by the inverse logit function:

\[
P_i = \frac{1}{1 + e^{-z}}
\]

(Fig. 1). The literature of anaesthesia is now replete with risk scores that offer a prediction versus an OR magnitude necessary for useful prognostic discrimination; OR must be much larger for clinical usefulness.32

**Table 2 Calculation of the odds ratio**

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Count of patients with an event</th>
<th>Count of patients without an event</th>
<th>Total counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Absent</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
</tbody>
</table>

The data is displayed as a 2 x 2 table. The odds ratio (OR) is the odds of an event in the group with the risk factor present \((a/b)\) divided by the odds of an event in the group with risk factor absent \((c/d)\). OR = \( a/b \). The standard error of the log OR is \( SE_{log(OR)} = \sqrt{1/a + 1/b + 1/c + 1/d} \). This is a univariable OR; only one risk factor is used in the estimation of the value. In the logistic model

\[
\log \left( \frac{P_i}{1 - P_i} \right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij},
\]

the ORs of multiple risk factors are estimated simultaneously; the adjusted OR of the \( j \)th risk factor is \( e^\beta_j \). The OR of a risk factor from a logistic model is denoted adjusted because its value mathematically controls for the effects of the other risk factors in the model. For both unadjusted and adjusted ORs, the 95% confidence interval is given by the following:

\[
e^{-0.65 \times SE_{log(OR)}} < e^{log(OR)} < e^{0.65 \times SE_{log(OR)}}
\]

**Validation**

All the methodological problems of determining risk factors also apply to risk scores and probabilities. Clearly, the predictions must be valid for the patients of the dataset used for the development (internal validation).35 Numerous problems may arise when users desire to transfer predictive algorithms from one period or one practice setting to a different time or another place.36 These can include the following: the statistical model may have been excessively optimistic in the choice and weighting of predictor variables within the original data set; in a new time or a new place other variables not relevant to the original model may become important; predictor variables may no longer be ‘predictive’; and the
Risk factors and risk prediction

Fig. 1

Logistic curve. This figure is a plot of the cumulative distribution function of the standard logistic probability distribution \( \text{probability} = \frac{1}{1 + e^{-z}} \) and is used to determine the predicted probability for any z score. As z varies from negative infinity to positive infinity, probability increases from 0 to 1. For the risk scores -2, 0 and 2, the predicted probability is 0.12, 0.5 and 0.88; this is demonstrated by the intersection of dotted lines on the logistic curve.

Functional relationship of predictor to outcome may have changed. Even when a risk factor has a true association with outcome, it has been empirically observed that the OR is usually inflated (too large).\(^{27}\) External validation is always necessary before any risk score with probabilistic predictions can be accepted. Examples of anaesthesia risk scores for which external validation has been reported include PONV and postoperative mortality.\(^{6,34,37}\)

In 2000, Eberhart et al.\(^ {37}\) reported an external validation study of three published risk scores for postoperative nausea and vomiting and postoperative vomiting in adults during the first 24 h following surgery. The assessment of predictors in an external validation study is illustrated with a re-analysis of these results in the Supplementary Digital Content 1 (http://links.lww.com/EJA/A23).

Overall performance measures
Frameworks for the analysis of external validation studies have been proposed.\(^ {38-41}\) The mathematical distance between the forecast and the actual outcome is the essence of quantifying overall performance – the shorter the distance, the better the model. Overall performance is measured by the distance of the predicted outcome \( (P_i) \) from the actual outcome \( (Y_i) \) in which \( Y_i \) is set to 0 or 1 for the non-occurrence or occurrence of the outcome. A good model of risk will have a short average distance. The
accepted measures for overall performance of the risk score in the validation datasets are the Brier score statistic and Nagelkerke's $R^2$ statistic. Nagelkerke's $R^2$, ranging from 0 to 1, is the proportion of the variation of the response variable explained by the risk score. In the Eberhart data, the predictor scores of Apfel et al.,

Koivuranta et al.,

and Palazzo and Evans

had Nagelkerke's $R^2$ values of 0.15, 0.19 and 0.09. Thus, only a small part of the variation of probability of PONV (15, 19 and 9%, respectively) among patients was explained by the $P_i$ calculated from the risk score.

**Discrimination**

Discrimination is the ability of the predictions to rank order patients with different outcomes. Discrimination of risk scores uses the notation of diagnostic tests (Table 3a and b)\(^9\) and is displayed by the receiver operating characteristic (ROC) curve. With the multiple values of $P_i$ from a risk score, multiple $2 \times 2$ tables may be calculated (one less than the number of distinct $P_i$ values), choosing all possible threshold values of $P_i$. An ROC curve is the line plot connecting the paired (sensitivity and specificity) values of these multiple $2 \times 2$ tables and extends the performance of a prediction across the prediction space. If all patients with outcome 1 have predictions $P_i$ greater than the largest prediction of the patients with outcome 0, then there is perfect discrimination. Conversely, if the predictions are uniformly interspaced for the patients with outcomes 1 and 0, then there is no discrimination.

The area under the curve (AUC) is the measure of discrimination; it is also called the concordance (C)-statistic (range 0–1). This is the probability that given two patients, one with and one without the eventual occurrence of an event, a larger risk score will be assigned to the former.\(^{45}\) Perfect discrimination has a C-statistic of unity, whereas a C-statistic of 0.5 indicates random predictions (no better than flipping a coin); a C-statistic of 0.5 is the area under the diagonal line from lower left to upper right. The C-statistic compares the ranking of $P_i$ scores in the study by Eberhart et al.\(^{42,44}\) had very similar AUCs of about 0.7 (Fig. 2). Another aspect of discrimination is the difference in the mean of predictions between outcomes; this is denoted the discrimination slope.\(^{39}\) For the three risk scores, the mean $P_i$ for those without and with an event were Apfel et al. (0.21, 0.32; slope = 0.11), Koivuranta et al. (0.34, 0.48; slope = 0.14) and Palazzo and Evans (0.14, 0.26; slope = 0.12). Although the absolute $P_i$ were different, the discrimination slopes were approximately the same, as were the AUCs.

**Calibration**

Discrimination and calibration are distinct properties of risk scores. Calibration is the agreement between observed outcomes and predictions (i.e. the correctness of prediction probabilities on an absolute scale). For example, if the predicted probability for in-hospital mortality is 20%, then about 20% of the patients with that predicted probability should die in the hospital. Calibration is assessed by regression statistics and is visualised by a calibration graph. These include a goodness-of-fit $\chi^2$-statistic by Hosmer–Lemeshow and a logistic regression by Cox.\(^{48,49}\) The plot for one of the risk scores in the study by Eberhart et al. shows poor calibration with both over-prediction and under-prediction (Fig. 3).\(^{37,44}\)

**Discussion**

With the advent of statistical software and data collection tools, hundreds of studies of prognosis are published annually, commonly involving the fate of patients with cancer, heart disease, pulmonary disease, trauma and so

---

**Table 3** The $2 \times 2$ tables diagnostic test definitions and an example

<table>
<thead>
<tr>
<th>(a) Diagnostic test definitions</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Prediction probability $&gt; P_i$</td>
<td>TP</td>
</tr>
<tr>
<td></td>
<td>FN</td>
</tr>
<tr>
<td>Sensitivity = TP/(TP + FN)</td>
<td>Specificity = TN/(FP + TN)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) Example</th>
<th>Postoperative vomiting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction probability $&gt; 0.275$</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>170</td>
</tr>
<tr>
<td>Sensitivity = 69%</td>
<td>Specificity = 69%</td>
</tr>
</tbody>
</table>

In this external validation sample, a threshold $P_i$ of 0.275 for the Apfel Score had a PPV of 31%; FN, false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; TN, true negative; TP, true positive.
on. We have highlighted methods for the development and interpretation of risk scores. Apart from studies of mortality, the anaesthesia literature also includes reports of scoring systems for the prognosis of everyday problems such as airway management, postoperative shivering, postoperative nausea, postoperative cognition and so on. One of the goals is to obtain risk probabilities adjusted for specific patient characteristics. However, these studies sometimes do not complete some of the necessary steps for prognosis research. In their qualitative systematic review, Ip et al. identified 48 studies reporting predictors of postoperative pain and analgesic consumption. Only two studies reported any validation of their predictors; neither study had an external validation.

Wyatt and Altman posed this provocative question more than a decade ago: ‘prognostic models, clinically useful or quickly forgotten?’ Inevitably it seems the first publication of a forecast system is excessively optimistic,
as later use is accompanied by a degradation of the predictions. Generalisability is the ability of the forecast to provide accurate predictions in a new sample of patients. The forecast should, of course, be reproducible in patients from the identical population obtained contemporaneously at the original institution; this is an internal validation. More importantly, the forecast should be transferable to different populations; external validation is necessary to show that temporal changes (a later year), geographic changes (a different continent), a different spectrum of illness and so on which do not defeat the prediction rule. What if external validation shows unfavourable characteristics of a scoring system? Several choices are evident and are as follows. First, among the three prediction scores for postoperative nausea and vomiting, that proposed by Koivuranta et al. was better when assessed by Eberhart et al. Yet, at another external site a different risk score may be the best. It may be decided that scoring system predictions are so inconsistent when applied elsewhere that they should be abandoned. Second, a prediction system for severe postoperative pain was derived and then tested by its originators at a different hospital several years after the initial publication. Statistical methods were used to recalibrate the prediction rule. It is unresolved if adjusting clinical prediction rules to fit a specific institution will prove practical and useful. Third, changes in medical care and patient risk may eventually degrade the

---

**Fig. 3**

Regression calibration plot for postoperative nausea and vomiting. The prediction probabilities for postoperative nausea and vomiting proposed by Palazzo and Evans were assessed by Eberhart et al. in 1,444 patients. The solid curved line reflects the observed calibration of outcomes to predictions (the 45° diagonal dashed line of identical predicted and actual probabilities represents ideal calibration). Below a predicted probability of 0.6, postoperative nausea and vomiting was under-predicted; above a predicted probability of 0.6, postoperative nausea and vomiting was over-predicted. There is poor calibration of predictions to outcomes.
calibration of even well established risk scores; 10 years after being published, the EuroSCORE predictor for cardiac surgery now over-estimates mortality and is being updated from a new large dataset.\textsuperscript{56,57} Wyatt and Altman\textsuperscript{51} also insisted that evidence of clinical effectiveness should be expected for risk scores and risk prediction. Risk estimations can improve a patient’s informed consent, are useful in benchmarking between institutions and physicians and may help in resource allocation. However, can the clinician use the presence of risk factors or a risk score as decision aids to change care and change outcomes?\textsuperscript{58} Even with the publication by an international panel of guidelines for PONV management, including the use of risk factors to guide therapy, there is still controversy about the usefulness of PONV risk scores.\textsuperscript{3,25,59,60} Anaesthesiologists should demand empirical evidence from randomised controlled trials that a risk score is clinically effective.\textsuperscript{51,61} In fact, our risk predictions might be reconfigured to be conditioned on the chosen anaesthetic management – the treatment choices to be included in our risk score calculations along with biomarkers and clinical risk factors.\textsuperscript{61} Risk factors and risk scores will remain a subject of intense interest and research.

Acknowledgements
No external funding or conflicts of interest declared.

References
2 Casey BM, McIntyre DD, Leveno KJ. The continuing value of the Apgar score for the assessment of newborn infants.\textit{N Engl J Med} 2001; \textbf{344}:467–471.
3 Oakland M. Grading of patients for surgical procedures.\textit{Anesthesiology} 1941; \textbf{2}:281–284.
9 Summary of the National Halothane Study. Possible association between halothane anesthesia and postoperative hepatic necrosis.\textit{JAMA} 1966; \textbf{197}:775–788.
11 Pace NL. Independent predictors from stepwise logistic regression may be nothing more than publishable p values.\textit{Anesth Analg} 2008; \textbf{107}:1775–1778.
14 Ellenberger C, Taït G, Beattie WS. Chronic beta blockade is associated with a better outcome after elective noncardiac surgery than acute beta blockade: a single-center propensity-matched cohort study.\textit{Anesthesiology} 2011; \textbf{114}:817–823.
20 Brotman DJ, Walker E, Lauer MS, O’Brien RG. In search of fewer independent risk factors.\textit{Arch Intern Med} 2005; \textbf{165}:138–145.
26 Levinson SS. Weak associations between prognostic biomarkers and disease in preliminary studies illustrates the breach between statistical significance and diagnostic discrimination.\textit{Clin Chim Acta} 2010; \textbf{411}:467–473.
27 Ioannidis JP. Why most discovered true associations are inflated.\textit{Epidemiology} 2008; \textbf{19}:640–648.
28 Ray P, Le Manach Y, Rou B, Houé TT. Statistical evaluation of a biomarker.\textit{Anesthesiology} 2010; \textbf{112}:1023–1040.
38 Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.

Risk factors and risk prediction


45 Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic curve (ROC). Radiology 1982; 143:29–36.


47 Qin GS, Hotilovac L. Comparisons of nonparametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Stat Methods Med Res 2008; 17:207–221.


52 Eberhart LHJ, Morin AM. Risk scores for predicting postoperative nausea and vomiting are clinically useful tools and should be used in every patient: con – ‘life is really simple, but we insist on making it complicated’. Eur J Anaesthesiol 2011; 28:155–159.


59 Pierre S. Risk scores for predicting postoperative nausea and vomiting are clinically useful tools and should be used in every patient: ‘don’t throw the baby out with the bathwater’. Eur J Anaesthesiol 2011; 28:160–163.
