Understanding the evidence: risk scores and risk prediction*

Nathan L Pace
Dept, of Anesthesiology, University of Utah, Salt Lake City, Utah, USA

Saturday, 11 June 2011 13:00 - 13:45 Room: E102

* This Refresher Course lecture will also be considered for publication in the European Journal of Anaesthesiology

Anaesthesiologists: pioneers of risk assessment and prognosis

Virginia Apgar was an experienced anaesthesiologist and an astute clinician. Based on her careful observations of thousands of newborns, she proposed in 1953 a 'New Method of Evaluation' that became the Apgar Score (five signs summing up to a score of 10) [1]. Infants had a mortality proportional to their Apgar Score - 14% (Apgar Score 0 to 2), 1.1% (Apgar Score 3 to 7), and 0.1% (Apgar Score 8 to 10). Fifty years later the Apgar Score is just as relevant in predicting neonatal mortality - 24% (Apgar Score 0 to 3), 0.9% (Apgar Score 4 to 6), and 0.02% (Apgar Score 7 to 10) [2]. Even earlier, in 1941, the American Society of Anesthesiologists promulgated a physical status score (the ASA score) of patients prior to anaesthesia and surgery [3]. While 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].

One or more diagnoses, that is - the surgical indication, have brought patients to us; other concomitant diagnoses may be present. Pre-operatively, our patients want us to tell them what will happen. "Will I survive the anaesthesia and surgery?" "Will I be sick?" "How severe will be my pain?" This is the human longing to predict the future. Prognosis is to know before, to give a forecast of the probable course and outcome of a disease, a procedure, a drug, etc.

Diagnosis, prognosis and risk

Diagnosis reflects the current vitality of the patient, their present condition. A numerical Apgar Score is a diagnosis of neonatal vigor; from low to high, it triggers a spectrum of resuscitation from aggressive efforts to merely observation. As previously noted, the Apgar Score is also a risk score for prognosis.

The baseline risk of an outcome for a population can come from several sources including: 1. large, randomised controlled trials; 2. observational studies; and 3. 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 peri-operative mortality was by a review of medical records at ten US university hospitals. Records of 600 000 patients during 5 years (1948 to 1952) uncovered 8 000 deaths for an in-hospital mortality rate of 1.3%, or one death for each 75 surgical procedures; of these, expert opinion classified only 224 deaths as anaesthetic related [5]. Fifty years later a US Medicare administrative database (MEDPAR) with 5 250 000 surgical patients from the years 2001 to 2006 had an in-hospital mortality of 3.1%, or 1 death for each 32 surgical procedures [6]. It is unlikely that this higher current mortality reflects a deterioration of care, but rather an aging population, a different mix of surgical procedures, an increased prevalence of concomitant diseases, the redirection of healthy patients to ambulatory surgery, etc. Clearly, some risk probabilities should be deprecated as reflecting earlier or different times.

The extreme outcomes of surgery are life or death; this is one meaning of risk - the listing of possible outcomes. But the forecast, the prognosis, is always a probability issue - probability (P) being a real number between 0 and 1. Thus the relevant definition of risk is the probability that an event will occur. Probability reflects the continuum between absolute certainty (P = 1) and certified impossibility (P = 0). We are unable to foretell the future of our patients individually, but we want to inform them of the probable outcome for a group of similar patients. (This is known as the frequentist view of probability - the relative frequency of an event 'in the long run' observations of many patients). Establishing probabilities is not the same thing as foretelling the future for an individual patient.
Using the MEDPAR data, an older patient could be informed pre-operatively that their $P_i$ is 0.031, i.e. 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.

**Risk factors**

**Effect size**

In statistics, an effect size is a measure of the strength of the relationship between two variables in sample data. The effect size is a descriptive statistic that numerically defines the estimated magnitude; it is a point estimate - the best guess as to effect in the underlying population. To make inferences about the effect size requires statistical tests and equivalently the calculation of 95% Confidence Intervals (CI). The magnitude of risk factors is usually specified by either the odds ratio (OR) or the risk ratio (RR). In an unadjusted OR, only two variables, the binary outcome and the binary predictor, are used for the calculation in a simple algebraic calculation (Table 1). The calculation of the CI is also straightforward. There are often multiple patient descriptors or covariates (age, sex, blood pressure, disease status, etc.) that are risk predictors. An adjusted OR is a statistical method for isolating the independent effect of covariates. The statistical methods for adjusted ORs are considerably more complicated and require iterative solutions to multivariable logistic regression equations [7]. Effect sizes from individual studies may be combined to give a summary effect size in a meta-analysis. The OR of a non-significant risk factor is centered at the value of unity (OR = 1) with the upper and lower bounds of the 95% CI being on either side of the line of identity (OR = 1).

**Table 1**

**Method for calculation of the Odds Ratio (OR)**

The data from a study can be redisplayed as a two by two table where $a$, $b$, $c$, and $d$ are the count of patients in each cell:

<table>
<thead>
<tr>
<th></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>New treatment</td>
<td>$a$</td>
<td>$b$</td>
<td>$a + b$</td>
</tr>
<tr>
<td>Other treatment</td>
<td>$c$</td>
<td>$d$</td>
<td>$c + d$</td>
</tr>
</tbody>
</table>

The OR is the odds of an event in the new treatment group ($a/b$) divided by the odds of an event in the other treatment group ($c/d$), i.e. $OR = ad/bc$. The precision of the OR is given by a 95% Confidence Interval (CI) which is derived from the standard error. The standard error of the log OR is:

$$SE_{\log(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

The 95% CI for the OR is given by:

$$e^{\left(\log(OR) - 1.96 \times SE_{\log(OR)}\right)} \text{ to } e^{\left(\log(OR) + 1.96 \times SE_{\log(OR)}\right)}$$
Biomarkers

A biomarker is an endogenous substance used as an indicator of a biological state. More specifically, a biomarker indicates a change in the expression, concentration or state of a protein that correlates with the risk or progression of a disease. Biomarkers are chosen because of their presence in the causal pathway of outcomes. To show an association between a biomarker and an outcome, biomarker plasma concentrations are first dichotomized by a threshold value as being positive or negative. Next, the odds of the binary outcome when the biomarker is positive is divided by odds of the outcome when the biomarker is negative and expressed as the OR. Some anaesthesia research in prognosis has focused on the biomarker brain natriuretic peptide (BNP) as a risk factor for major adverse cardiovascular events (MACE: cardiac death, non-fatal myocardial infarction) after non-cardiac surgery. BNP, a circulating hormone synthesised by cardiomyocytes in response to increased ventricular wall stress or ischaemia, has natriuretic and vasodilator properties; BNP is a risk factor for heart failure. In a systematic review of 15 studies (about 5 000 patients having non-cardiac surgery), an elevated pre-operative BNP had a highly elevated OR (19.77; 95% CI 13.18-29.65) for short-term MACE [8]. Interpreting this very large OR was made difficult by the lack of a common definition for the threshold value to separate BNP into normal and abnormal ranges, a general problem in systematic reviews of prognosis [9]. The generally accepted range of normal plasma BNP is up to 100 pg/ml; in the 15 studies, the threshold separating normal BNP from abnormal BNP ranged widely from as low as 35 pg/ml to as high as 255 pg/ml.

Clinical risk factors

Observational data has become a fertile source of material for epidemiological reports; these patient records have been collected on diseases, treatments, events, outcomes, etc. by hospital and government information systems. Virginia Apgar used clinical judgment to derive the Apgar Score. Current practice for identifying risk factors is very different. About 50 years ago an algorithm was proposed for an automatic procedure to select a statistical model (i.e. choose risk factors) where 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 algorithm has been incorporated into regression analysis for both linear and logistic models and is implemented variously in most statistical software packages. There are many variations of automatic variable selection. Most research reports use this automatic procedure even when the authors had prior knowledge about likely risk factors.

There are widespread methodological shortcomings in the primary studies of prognosis. These potential biases have been enumerated and include failure to clearly define and describe the source population and to adequately measure the putative prognostic factors [9]. Another weakness is the inconsistent analysis and reporting of the potential risk factors. Ip et al found 48 studies (23 037 patients) reporting risk factors for postoperative pain and analgesic consumption [11]. Pre-existing pain, anxiety, age and surgery type were the four most important risk factors for the intensity of postoperative pain. Yet of the 32 of 48 studies reporting on postoperative pain intensity, only eight, 15, 12, and six reported an analysis of these risk factors. Were these risk factors not analysed in the primary studies? Were these risk factors actually analysed in the other 24, 17, 20, and 26 studies, but not reported because no association with pain intensity was found? The answer is not known.

Multivariable probabilistic risk predictions

Estimation

A probability prediction rule assigns a probability to a patient for the occurrence of a specified event; these assigned probabilities are variously called a prediction, a forecast, or a prognosis - nomenclature often used interchangeably. These predictions are mainly descriptive, not mechanistic or explanatory, of the outcome event. The raw materials for developing a probability prediction rule are covariates (e.g. sex, ASA physical status, surgery type, age, sodium concentrations) recorded prior to the occurrence of the outcome event and the occurrence or nonoccurrence of the event.
Using essentially the same mathematical methods (stepwise logistic regression) as for identifying risk factors, a statistical model is estimated from the training data set. For example, in a logistic model the probability of the event ($\pi$) is related to a linear combination of the covariates ($x_1, \ldots, x_k$) by the logit link function:

$$\ln\left(\frac{\pi}{1-\pi}\right) = z = \sum_{j=0}^{k} \beta_j x_j = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

This model, that is the covariates chosen and the regression (parameter) coefficients estimated ($\beta_0$, $\beta_1$, …, $\beta_k$), is selected to maximize the goodness of fit of covariates values to the presence/absence of the event. These covariates are called independent predictors. The ‘$z$’ is the risk score for the model. Inherent in the model will be a formula or algorithm for the calculation of a specific prediction probability value $P$, for a new patient with their set of covariate values. This is calculated by the inverse logit function:

$$P = \frac{1}{1 + e^{-z}}$$

The literature of anaesthesia is now replete with risk scores that offer a $P$ formula. For example, $z = -2.28 + 1.27 \times$ (female sex) + 0.65 \times (history of PONV or motion sickness) + 0.72 \times (non smoker) + 0.78 \times (postoperative opioid administration)$ is a risk score for PONV. A non-smoking female with a history of motion sickness receiving postoperative morphine would have a calculated risk score of 1.14 with a $P$ of 0.76 - a 76% probability of having PONV [12, 13].

Unfortunately, some of the consequences for using stepwise logistic regression include over optimistic (too large) regression coefficients, spuriously narrow confidence intervals on coefficients and predictions, and the inclusion of noise (unrelated covariates) in the model. In one statistical simulation, randomly generated covariates and outcomes were used to check if stepwise logistic regression finds spurious associations. Disturbingly this method created in over 80% of the simulations the appearance of an association between a random binary outcome and one or more randomly created explanatory variables. From noise, independent predictors with publishable p values less than 0.05 had been created [10].

Validation

Numerous problems may arise when users desire to transport predictive algorithms from one period or one practice setting to a different time or another place. These can include: 1. the statistical model may have been excessively optimistic in the choice and weighting of predictor variables within the original data set; 2. in a new time or a new place other variables not relevant to the original model may become important; 3. predictor variables may no longer be ‘predictive’; and 4. the functional relationship of predictor to outcome may have changed. Empirical experience has demonstrated that temporal transportability of a predictor (subsequent patients at the institution developing the predictor) by itself does not ensure geographical transportability [14]. External validation is always necessary before any risk score with probabilistic predictions should be accepted. Examples of external validation of anaesthesia risk scores include PONV [12, 13] and postoperative mortality [6].

A framework for the analysis of external validation studies has been proposed [14]. Overall performance is measured by the distance of the predicted outcome ($P$) from the actual outcome ($Y_i$); a good model of risk will have a short average distance. The accepted measures for overall performance in the validation datasets are the Brier score statistic and Nagelkerke’s R2 statistic. The overall performance can be partitioned into two characteristics - discrimination and calibration. Statistical software tools for estimation of overall performance, discrimination and calibration are readily available.

The $c$ statistic is a measure of discrimination; it is a rank order statistic for predictions versus actual outcomes and is equivalent to the area under the receiver operating characteristic curve (AUC). Perfect discrimination corresponds to a $c$ statistic of 1 and is achieved if the $P$ for all patients with an event is higher than those for all patients without an event, with no overlap. A $c$ statistic value of 0.5 indicates a
risk score without discrimination, i.e. no better than flipping a coin. While a good risk model will have a high discrimination, by itself the c statistic is not optimal in assessing or comparing risk models [15].

The third aspect of performance measures is calibration, i.e. the agreement between observed outcomes and predictions. 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.

**Risk predictions and treatment**

Wyatt and Altman posed this provocative question a decade ago: ‘Prognostic models, clinically useful or quickly forgotten?’ [16]. This is a relevant question for biomarkers, risk factors, and risk scores with probabilistic predictions. 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 transportable 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, etc, do not defeat the prediction rule.

Wyatt and Altman also insisted that evidence of clinical effectiveness should be expected for risk scores and risk prediction [16]. Besides giving a forecast to the patient, can the clinician use the presence of risk factors or a risk score for mortality as decision aids to change care and change outcomes? To a large extent, this is unknown at present. In fact, our risk predictions should be conditioned on the anaesthetic management to be chosen; i.e. treatment choices should be included in our risk score calculations along with biomarkers and clinical risk factors [17]. Even with accurate risk scores, anaesthesiologists should demand empirical evidence from well designed and accomplished clinical trials that the model is clinically effective. These trials should be controlled studies in which the effects of providing predictions on actual clinical practices or patient outcomes are measured. Proof of effectiveness will require the methodology of randomized controlled trials [16, 17].

**Key learning points**

- Risk is a probability issue - probability (P) being a real number between 0 and 1. The definition of risk is the probability that an event will occur. Probability reflects the continuum between absolute certainty (P = 1) and certified impossibility (P = 0).
- Baseline risk probabilities for mortality are obtained from large observational studies and administrative databases.
- Risk factors include biomarkers and clinical covariates.
- Risk scores are usually obtained by stepwise logistic regression with the inverse logit function used to calculate the probability of the event in a future patient. Risk scores should be externally validated by patient outcomes at other locations and at other times.
- The properties of risk scores require assessment by measures of overall accuracy, discrimination, and calibration.
- Evidence from randomized, controlled clinical trials should be obtained to show the effectiveness of risk scores for patient management.
References