Can J Anesth/J Can Anesth (2016) 63:148–158
DOI 10.1007/s12630-015-0537-2

REVIEW ARTICLE/BRIEF REVIEW

Predicting outcomes: Is there utility in risk scores?
La prévision des pronostics: les cotes de risque sont-elles utiles?

Duminda N. Wijeysundera, MD, PhD

Received: 9 June 2015 / Revised: 19 October 2015 / Accepted: 11 November 2015 / Published online: 15 December 2015

Abstract

Purpose *This review discusses the utility of risk scores, specifically, the role of preoperative risk scores in guiding the management of surgical patients, approaches to evaluate the quality of risk scores, and limitations to consider when applying risk scores in clinical practice.*

Principal findings *This review shows how accurate predictions of perioperative risk can help inform patients and clinicians with respect to decision-making around surgery; identify patients who warrant further specialized investigations, new interventions intended to decrease risk, modifications in planned operative procedures, or intensification of postoperative monitoring; and facilitate fairer comparisons of outcomes between providers and hospitals. A preoperative risk score formally integrates several pieces of clinical information (e.g., age, comorbid disease, laboratory tests) to arrive at an overall estimate of an individual patient's expected risk for specific postoperative adverse events. A good risk score should be simple to incorporate in clinical practice, reliable when applied by different raters, and accurate at predicting postoperative risk. Several analytical methods (e.g., receiver operating characteristic curves, likelihood ratios, risk reclassification tables, observed vs predicted plots) are required to characterize the relevant domains that encompass the prognostic accuracy of a risk score. External validation is critical in determining whether the predictive accuracy of a risk score is preserved when applied to new settings, populations, or outcome events.*

Conclusions *Preoperative risk scores help inform perioperative clinical decision-making. Future research must determine how estimates of preoperative risk can be updated with information from the intraoperative period, how risk information should be communicated to patients, and which interventions can improve outcomes among patients within newly identified risk strata.*

Résumé

Objectif *Ce compte rendu s'intéresse à l'utilité des cotes de risque, et plus spécifiquement au rôle des cotes de risque préopératoires dans l'orientation de la prise en charge des patients chirurgicaux, aux approches permettant d'évaluer la qualité des cotes de risque, ainsi qu'aux limites à garder à l'esprit lorsqu'on applique de telles cotes de risque à la pratique clinique.*

Constatations principales *Ce compte rendu démontre comment des prédictions précises du risque périopératoire peuvent aider à guider les patients et les cliniciens dans leurs prises de décision concernant la chirurgie, à identifier les patients qui bénéficieraient d'examens plus spécialisés, de nouvelles interventions visant à réduire leur risque, de modifications dans les interventions opératoires planifiées, ou d'une intensification du monitorage*

D. N. Wijeysundera, MD, PhD
Department of Anesthesia, Toronto General Hospital and University of Toronto, Toronto, ON, Canada

D. N. Wijeysundera, MD, PhD
Li Ka Shing Knowledge Institute of St. Michael's Hospital, Toronto, ON, Canada

D. N. Wijeysundera, MD, PhD
Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, ON, Canada

D. N. Wijeysundera, MD, PhD (✉)
Department of Anesthesia and Pain Management, Toronto General Hospital, Eaton Wing 3-450, 200 Elizabeth Street, Toronto, ON M5G 2C4, Canada
e-mail: d.wijeysundera@utoronto.ca

postopératoire, et à faciliter des comparaisons plus justes des pronostics entre différents fournisseurs et hôpitaux. Une cote de risque préopératoire intègre de façon formelle plusieurs éléments d'informations cliniques (par ex., l'âge, les maladies comorbides, les tests de laboratoire) afin d'obtenir une estimation globale du risque attendu, pour un patient en particulier, de subir des complications postopératoires spécifiques. Pour être utile, une cote de risque doit être simple à intégrer dans la pratique clinique, fiable lorsqu'elle est appliquée par différents évaluateurs, et précise dans sa prédiction du risque postopératoire. Plusieurs méthodes analytiques (par ex., les courbes ROC, les rapports de vraisemblance, les tables de reclassification des risques, les représentations graphiques des observations vs des prévisions) sont nécessaires afin d'identifier les domaines pertinents qui contribuent à la précision pronostique d'une cote de risque. La validation externe est essentielle pour déterminer si la précision prédictive d'une cote de risque est conservée lorsqu'on l'applique à un nouveau cadre, à d'autres populations ou à d'autres événements pronostiques.

Conclusion *Les cotes de risque préopératoires constituent un outil utile pour informer la prise de décision clinique périopératoire. Les recherches futures doivent déterminer comment mettre à jour les estimations du risque préopératoire en se servant d'informations tirées de la période peropératoire, comment les informations de risque devraient être communiquées aux patients, et quelles interventions peuvent améliorer les pronostics parmi les patients inclus dans les strates de risque nouvellement identifiées.*

Prophesy is a good line of business, but it is full of risks.
—Mark Twain

An integral part of the practice of anesthesiology and perioperative medicine is the assessment of surgical patients' risks for future adverse outcomes. Indeed, every preoperative assessment or consultation involves, either explicitly or implicitly, an estimation of the individual patient's risks for major adverse postoperative outcomes, such as death or major complications.

What are the potential roles for information on estimated perioperative risk?

Why would an accurate estimate of perioperative risk be important for clinical care? *First*, it helps better inform patients' and clinicians' decision-making surrounding surgery. Indeed, for patients considering surgery, a critical component of the informed consent process is consideration of the potential risks of the planned procedure. In some cases, patients might refuse the original planned surgery or consider less-invasive alternatives when informed that the predicted perioperative risk is very high. *Second*, predictions about perioperative risk can help determine the need for further specialized investigations such preoperative pulmonary function testing or cardiac stress testing. For example, clinical practice guidelines from the American Heart Association and American College of Cardiology recommend that preoperative stress testing be considered only if an individual's expected risk of postoperative cardiac complications exceeds a minimum threshold.[1] *Third*, information on risk can help determine which patients might benefit from specific interventions. For example, prior observational studies have suggested that perioperative beta blockers are associated with benefit in patients at higher predicted cardiac risk, while they are associated with harm in individuals at lower predicted cardiac risk.[2,3] *Fourth*, risk information can help specify the required intensity of perioperative monitoring, such as whether a patient warrants invasive monitoring or postoperative care in a critical care unit. The role of estimated perioperative risk in rationally choosing postoperative care settings is especially important since monitored acute care beds are expensive and often scarce resources. Furthermore, the "failure-to-rescue" paradigm suggests that selective improved postoperative monitoring of high-risk patients might be a critical avenue for improving overall postoperative outcomes. Specifically, comparisons of postoperative outcomes across acute care hospitals in the United States found that patients at better performing hospitals do not necessarily experience lower rates of complications but rather lower complication-associated mortality rates.[4] These data point to the importance of managing surgical patients at increased risk for complications in postoperative environments that facilitate earlier detection and treatment of those complications.

Information of patients' estimated risks for adverse postoperative outcomes can also be useful in contexts outside clinical care. This information is vital to any comparison of performance across physicians and hospitals,[5,6] such as whether postoperative outcomes are better or worse at a specific hospital when compared with other healthcare facilities. Since some hospitals are more likely to care for sicker patients, fair comparisons rely on statistical risk adjustments for important differences in surgical case mix. These statistical methods for risk adjustment, in turn, require accurate estimates of expected postoperative risk. Estimates of patients' expected risks for adverse postoperative events are also helpful in the design of research studies, such as informing the selection of participants for randomized-controlled trials or prospective cohort studies. For example,

previously identified predictors of postoperative pulmonary complications were used to design the inclusion criteria for a recent randomized-controlled trial of protective mechanical ventilation in high-risk patients undergoing major abdominal surgery.[7,8]

Simple methods to assess perioperative risk

A perioperative physician evaluating a patient awaiting surgery should be able to make an initial judgement about their expected perioperative risks using readily available clinical information without the need for additional specialized investigations. Previous research has shown that several broad classes of preoperative information can contribute to this overall estimate of risk, including patients' demographics (e.g., age, sex); presence, burden, and severity of comorbid diseases (e.g., coronary artery disease, heart failure, chronic obstructive pulmonary disease); complexity and urgency of the planned surgical procedure; patients' functional capacity; and selected conventional laboratory test results (e.g., hemoglobin, estimated glomerular filtration rate).[9-11] A critical challenge that physicians face is how best to *integrate* these various sources of information to arrive at an overall estimate of risk for an individual patient.

Importantly, one of the most common approaches for estimating perioperative risk relies on a physician's subjective evaluation of a patient's overall health status, namely, the American Society of Anesthesiologist's Physical Status (ASA-PS) scale.[12] A potential limitation that should be considered with any subjective assessment of risk is the potential for significant inter-rater variability. For example, would ten different anesthesiologists assign an individual patient the same ASA-PS rating? Considerable inter-rater variation can reduce the accuracy of any predictive tool.[13] Notably, some prior research has found limited inter-rater reliability when anesthesiologists applied the ASA-PS scale to hypothetical case scenarios or de-identified medical records.[14-19] Conversely, more recent research has shown the scale to have at least moderate inter-rater reliability in usual clinical practice, with more than 98% of paired ASA-PS ratings of individual patients being within one class of each other.[20] Additionally, despite these potential limitations, the ASA-PS scale has shown at least moderate accuracy in predicting postoperative mortality across a wide range of studies.[21]

Estimating perioperative risk using risk scores

An alternative to a subjective evaluation of overall risk is the formal integration of several different sources of risk information into a single score, such that patients are allocated points based on the presence of prognostically important risk criteria (e.g., increased age, concomitant coronary artery disease, preoperative anemia). Over the past few decades, an increasing number of such risk scores have been developed to help predict specific types of outcomes in surgical patients. A recent systematic review identified 27 studies, published during 1980-2011, that evaluated 34 different risk stratification tools for predicting morbidity and mortality in major surgery.[22] Importantly, this review also excluded up to 1,100 other studies of preoperative risk scores because they focused on pediatric surgical patients, specific surgical subgroups (e.g., cardiac surgery, ambulatory surgery), or risk indices with insufficient validation. Stated otherwise, there is a very large body of literature on perioperative risk scores.

Some of these perioperative risk scores were initially developed for other purposes and were subsequently adapted for risk prediction in surgical patients. Some of these adapted scores are presented in Table 1. As an example, while the ASA-PS scale is now typically used to predict outcomes after surgery, it was originally developed to classify the preoperative health of surgical patients.[12] The Charlson Comorbidity Index, which has also been used to predict short-term postoperative outcomes in surgical patients, was originally developed to predict one-year mortality in medical inpatients.[23]

An alternative to adapting a previously developed scoring system is to develop a new index for the specific purpose of predicting outcomes after surgery. Selected examples of such risk scores are presented in Table 1. When developing a risk score *de novo*, the choice of variables to include and their relative weighting (i.e., how many points each different component should be allocated) are usually determined by prognostic modelling studies. In such a study, the first step would typically be to identify a group of surgical patients in whom a range of potentially important characteristics would be measured. These same individuals would then be followed to determine whether they develop the outcome of interest, such as death or postoperative respiratory failure. Statistical modelling techniques would be applied to determine which baseline characteristics were most predictive of an individual developing the outcome of interest. These statistical analyses help determine both the choice and relative weighing of variables to be included in the risk score. While a range of statistical methods can be applied to help identify the most predictive baseline characteristics, the most common approaches typically used are multivariable logistic regression modelling for dichotomous outcomes (e.g., 30-day postoperative myocardiaI infarction) or multivariable Cox proportional hazards modelling for time-to-event outcomes (e.g., survival over a two-year follow-up). Other less commonly used methods include

Table 1 Selected examples of perioperative risk scores

Risk Scores Specifically Developed for Risk Prediction in Surgical Patients		
Risk Score	Derivation Cohort	Outcome in Derivation Cohort
POSSUM[53]	General surgery	30-day death
P-POSSUM[40,41]	General surgery	30-day death
Revised Cardiac Risk Index[35]	Elective noncardiac surgery	In-hospital cardiac complications
Surgical Risk Scale[42]	General surgery	In-hospital death
Surgical Risk Score[54]	General surgery	In-hospital death
Surgical Apgar Score[51]	Colorectal surgery	30-day death
NSQIP Universal Risk Calculator[37]	Mixed surgical cohort	30-day death or complications
Mallampati score[55]	Mixed surgical cohort	Difficult tracheal intubation
Euroscore[56]	Cardiac surgery	30-day or in-hospital death
Cardiac Anesthesia Risk Evaluation Score[57]	Cardiac surgery	In-hospital death or morbidity
Pulmonary risk index of Arozullah *et al.*[7]	Noncardiac surgery	30-day postoperative pneumonia
Delirium risk index of Marcantonio *et al.*[58]	Noncardiac surgery	In-hospital postoperative delirium

Risk Scores Adapted from Other Settings and Purposes		
Risk Score	Derivation Cohort	Original Purpose
ASA-PS classification[12]	General surgery	Classifying preoperative health of surgical patients
APACHE II score[59]	Patients admitted to critical care units	Predicting in-hospital death
MELD score [60]	Patients undergoing TIPS procedures	Predicting three-month mortality
Charlson Comorbidity Index[23]	Medical inpatients	Predicting death at one year

APACHE = Acute Physiology and Chronic Health Evaluation; ASA-PS = American Society of Anesthesiologists Physical Status; MELD = model for end-stage liver disease; NSQIP = National Surgical Improvement Quality Program; POSSUM = Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity; P-POSSUM = Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity; TIPS = transjugular intrahepatic portosystemic shunt

recursive partitioning[24] as well as artificial neural networks and other machine learning techniques.[25] Each analytic approach has specific advantages, limitations, and considerations (e.g., underlying model assumptions). Detailed discussions of these analytical techniques are available to readers in several available comprehensive textbooks on prognostic modelling.[26-28] In addition, the recently published "Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD)" statement provides readers with a checklist of important items that should be included in any published report of a prognostic modelling study.[29]

Prognostic performance: discrimination, risk reclassification, calibration, and validation

At the minimum, a good risk score must accurately predict risk. Assessing prognostic accuracy is not necessarily straightforward since it cannot be summarized with a single simple metric.[13] Instead, several characteristics must be considered. First, the quality of risk prediction can be evaluated based on *discrimination*, which describes the extent to which a risk score assigns different predicted risk estimates to individuals who did or did not have the outcome of interest. A good risk score should be more likely to assign a higher predicted risk to an individual who had an event than to one who did not. Thus, discrimination describes the degree of overlap in risk scores between individuals who do or do not develop the outcome of interest (Fig. 1). Smaller degrees of overlap indicate better discrimination. A very commonly used measure of discrimination is the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The AUC is related to other measures of discrimination commonly used in the diagnostic test literature, namely, sensitivity, specificity, positive predictive value, and negative predictive value. The AUC values range from 0-1 and measure the average probability that a risk score will assign an individual with an outcome event a higher predicted risk than an individual without an event. Thus, a score that performs no better than chance (i.e., flipping a coin) will have an AUC value of 0.5, while a score that perfectly separates individuals with and without events will have an AUC value of 1.

While good preoperative risk indices typically have AUC values of 0.75 or higher, it is challenging to translate information on AUCs to clinical decision-making. This

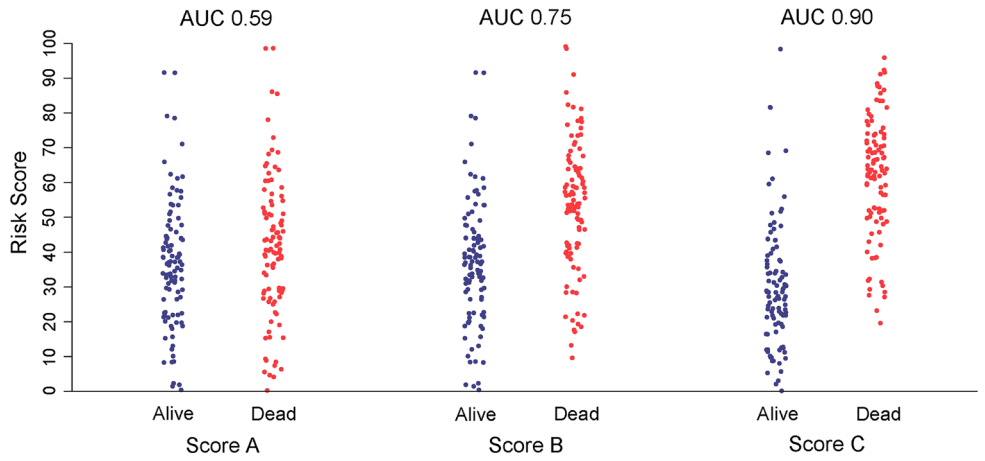

Fig. 1 Assessment of discrimination. Discrimination refers to the extent to which a risk score assigns different predicted risk estimates to individuals who did or did not have the outcome of interest. Fig. 1 presents three hypothetical risk scores for predicting postoperative mortality. Each score assigns an individual patient a score from 0-100. The range of scores among 200 individuals who did or did not have the outcome of interest (i.e., death) is separately presented for the three scores in Fig. 1. Score A has the most overlap in scores between individuals who were dead *vs* alive after surgery; hence, it has the lowest discrimination and lowest area under the curve (AUC) of the receiver operating characteristic curve. Conversely, Score C has the least overlap and therefore the highest discrimination and AUC. The graph was plotted using the R Statistical Language Version 3.2.1 (Vienna, Austria)

same challenge exists when trying to use sensitivity and specificity to guide clinical decision-making. An approach that can make this information more translatable to clinical decision-making is converting AUCs into likelihood ratios. Likelihood ratios essentially communicate the extent to which a risk score changes patients' predicted risk for an outcome from their baseline risk. For example, if the average rate of myocardial infarction (MI) in a cohort is 3%, a test result with an associated likelihood ratio of 3.0 means that the patient's expected risk is now approximately 9% or three times higher.[A] Conversely, a likelihood of 0.2 means that a patient's predicted risk is now only one-fifth the average or 0.6%. Clinically useful risk scores can shift an individual's risk from the average by a meaningful degree. It is been suggested that likelihood ratios greater than 2 or less than 0.5 are needed to provide even minimal additional information.[30]

Second, *risk reclassification* tables can be used to assess the performance of prognostic tools.[31] These methods are typically used to evaluate whether including a new variable, such as a biomarker, to a risk score improves overall risk prediction. This approach first requires specifying clinically relevant strata of risk. These strata should refer to clinically relevant different categories of predicted risk (i.e., categories with different implications for clinical decision-making). For example, a recent prospective cohort study that evaluated the additional value of preoperative coronary computed tomographic angiography (CTA) to usual clinical risk factors for predicting postoperative MI or cardiac death or non-fatal MI considered three strata of expected rates of the primary outcome, namely, $< 5\%$, 5-15%, and $>15\%$.[32] Importantly, the results of reclassification methods are sensitive to the number and definition of these strata.[33] Reclassification table analyses then evaluate the extent to which the addition of a new variable to the risk score improves assignment of patients to these different predicted risk strata. Specifically, they determine the net number of individuals with events who are assigned to higher predicted risk strata as well as the net number of individuals without events who are assigned to lower predicted risk strata. While these analyses might be more difficult to interpret, they can also be more informative than simple comparisons of AUCs between different risk scores. For example, addition of a new variable to a risk score can sometimes result in only relatively small changes in AUCs but significantly improved risk reclassification. Thus, the AUC may be insensitive to important improvements in discrimination that can impact on clinical decision-making.[34] In addition, small improvements in AUCs might mask considerable *worsening* in risk reclassification. For example, in the recent cohort study of preoperative coronary CTA,[32] addition of preoperative imaging to usual clinical risk factors resulted in a small improvement in AUC (0.62 to 0.66). Nonetheless, reclassification tables showed that, among 955 patients evaluated with preoperative coronary CTA, considerably more individuals without events were incorrectly assigned to higher risk strata ($n = 98$) than individuals with events who were correctly assigned to

[A] This is an approximate estimate (exact value is 8.5%), since likelihood ratios technically entail multiplying odds as opposed to probabilities of events occurring.

Table 2 Example of a reclassification table analysis in a prospective cohort study of preoperative coronary computed tomographic angiography

Individuals without events

	Predicted Risk Based on Risk Factors and Coronary CTA		
Predicted Risk Based on Risk Factors Alone	<5%	5-15%	>15%
<5%	191	114	0
5-15%	47	453	37
>15%	0	10	29

Individuals with events

	Predicted Risk Based on Risk Factors and Coronary CTA		
Predicted Risk Based on Risk Factors Alone	<5%	5-15%	>15%
<5%	5	10	0
5-15%	0	41	7
>15%	0	1	10

Example of a risk reclassification table analysis from a prospective cohort study that evaluated whether preoperative coronary computed tomographic angiography (CTA) improved the prediction of postoperative cardiac death or myocardial infarction.[32] The predicted risks of the primary outcome are separately reported for individuals with the event *vs* individuals without the event. The tables present the numbers of individuals with differing types of agreement between the two risk prediction models (i.e., clinical risk factors alone *vs* clinical risk factors plus coronary CTA): lower predicted risk category with addition of CTA, similar predicted risk category in the two models, and higher predicted risk category with addition of CTA. Addition of coronary CTA is helpful if individuals without events are moved to lower predicted risk categories or if individuals with events are moved to higher predicted risk categories (i.e., green cells). Conversely, coronary CTA is unhelpful if individuals without events are moved to higher predicted risk categories or if individuals with events are moved to lower predicted risk categories (i.e., orange cells). Individuals in grey cells are assigned to the same risk category by both risk prediction models. The overall improvement in risk classification by the addition of CTA is conveyed by the difference in the number of individuals with helpful risk reclassification (i.e., green cells) *vs* the number with unhelpful reclassification (i.e., orange cells)

higher risk strata ($n = 17$). Thus, on balance, this study did not find preoperative coronary CTA to be helpful with respect to improving risk prediction, which was a finding not evident based on AUC analyses alone (Table 2).

Third, risk scores can be characterized with respect to *calibration*, which refers to how well observed outcome event rates agree with event rates predicted by the risk score. For example, if a risk score predicted that individuals with a specific risk score would experience an event rate of 10%, how well does the observed event rate agree with this prediction? This is a critically important issue, since clinical decision-making is often based on predicted event rates. While calibration can be assessed using a range of statistical tests, such as the Hosmer-Lemeshow statistic, often the simplest and most transparent approach is simply to present the observed event rates within risk strata defined by expected event rates (Fig. 2). This graphical presentation provides an overall evaluation of calibration and also points to specific contexts where calibration may be poor. For example, a risk score may consistently overestimate event rates in high-risk individuals, which is an important consideration if this expected risk information were to be used to help determine whether a patient should be offered surgery.

Fourth, *validation* of risk scores is needed to determine whether a risk score has stable prognostic accuracy across different patient samples. In general, the most optimistic estimate of prognostic accuracy is seen within the original cohort in which a risk score was derived. Even within the original study, more conservative estimates of prognostic performance can be obtained using internal validation techniques such as data splitting or bootstrap resampling.[26] Nonetheless, all risk scores should ideally undergo external validation in new populations external to the one in which it was developed. Calibration can certainly worsen when a risk score undergoes external validation. There are many potential reasons for this degradation in performance, including differences in population characteristics that are

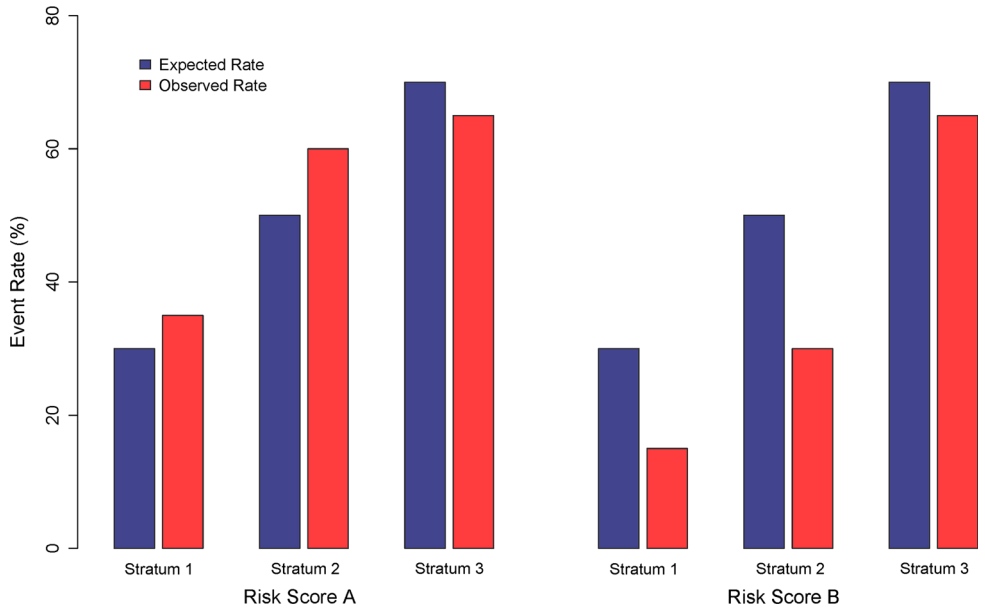

Fig. 2 Assessment of calibration. Calibration refers to how well the rates of observed outcome events agree with the rates of predicted events by a risk score. Fig. 2 presents the rates of observed (red columns) and predicted (blue columns) outcome events for two different hypothetical risk scores. Each risk score categorizes patients into one of three risk strata with differing expected risks. The red columns indicate that both risk scores perform reasonably well with respect to separating individuals with differing observed rates of outcome events. Nonetheless, when comparing the rates of expected *vs* observed outcome events, Risk Score A has better calibration such that the rates of observed and expected events were generally similar. By comparison, Risk Score B has poorer calibration with considerable overprediction of event rates in Strata 1 and 2. The graph was plotted using the R Statistical Language Version 3.2.1 (Vienna, Austria)

not accounted for by the risk score, temporal changes in overall outcomes (e.g., improvements in surgical outcomes over time), or differing methods for ascertaining outcomes. For example, many older predictive indices for estimating perioperative cardiac risk were derived in settings where postoperative MI was detected using creatine kinase MB assays,[35] whereas contemporary clinical practice typically relies on more sensitive troponin assays.[9] Thus, in all likelihood, the predicted MI rates in the original studies systematically underestimate rates observed in contemporary practice. Understanding the extent of this degradation in calibration is a critical reason why risk scores should undergo external validation before their widespread uptake into clinical practice.

Other characteristics needed in a good risk score

Aside from prognostic accuracy, what other characteristics are needed in a good preoperative risk score? Importantly, it should be simple and straightforward to implement into clinical practice.[13] Stated otherwise, a very accurate but complex risk score that requires computing weights and probabilities will likely be simply an academic exercise as opposed to a useful additional tool for the busy clinician. It is notable that the Revised Cardiac Risk Index (RCRI), one of the most commonly used preoperative cardiac risk indices, has only six components that are allocated one point each in a very simple weighting scheme.[35] While the RCRI does have moderate predictive accuracy,[35,36] its simplicity is arguably the major reason for its widespread uptake into clinical practice. A development that might allow for easier uptake of more complex risk estimation methods into clinical practice is the web-based risk calculator. A key example is the series of American College of Surgeons risk calculators (http://riskcalculator.facs.org). This web-based software implements a series of complex risk models to predict a range of major postoperative complications in a simple user-friendly interface. Software users must first input 22 readily available patient and surgery characteristics onto a web-based form. The software then executes the underlying complex calculations in the background[37] and ultimately feeds the user a report with estimated probabilities for a range of complications.

In addition, a good risk score should be reliable such that a single patient being assessed by several raters should be assigned very similar risk scores by said raters.[13] For example, the inconsistent accuracy of the Mallampati score in predicting difficult endotracheal intubation in external validation studies may be related[38] in part to inadequate inter-rater reliability.[39] As another example, the American College of Surgeons risk calculators incorporate the ASA-PS classification, which may also have uncertain inter-rater

reliability.[14-20] Thus, whenever possible, a risk score and its individual components should have good inter-rater reliability.

The bottom line: What is the utility of preoperative risk scores?

As indicated previously, there is now a very large body of literature on perioperative risk scores.[22] What is the clinical utility of these prediction tools? Among published tools, some have already been evaluated in multiple validation studies, with the Portsmouth Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (P-POSSUM) and the Surgical Risk Scale showing the most consistent accuracy in predicting mortality and morbidity after major surgery.[22,40-42] While current risk scores have limitations and can be further improved,[22] these data indicate that risk scores can indeed provide reasonably accurate estimates of perioperative risk and thereby *inform* perioperative care. Nonetheless, the critical question remaining is whether accurate estimation of risk can then translate into *improved outcomes*. While it is theoretically possible that improved targeting of care in both low-risk and high-risk patients can improve their outcomes, such a scenario requires that estimates of risk are accurate and that efficacious interventions are applied selectively in these different risk strata. At present, it remains uncertain whether improved preoperative estimation of risk does indeed translate into improved clinical outcomes after surgery. An important reason for this uncertainty is likely the paucity of interventions in perioperative medicine proven to improve postoperative outcomes in randomized-controlled trials.

Limitations of risk scores

What limitations of risk scores should be considered? This article has already discussed several potential limitations, including inadequate discrimination, reduced calibration with application of risk scores in new settings, and variable inter-rater reliability. There are also several other potential limitations that should be considered. *First*, risk scores are typically developed to predict a specific set of outcomes. The prognostic accuracy of a risk score is not necessarily transferable to other outcomes. For example, the RCRI was developed to predict perioperative cardiac complications.[35] In validation studies, it retained moderate discrimination when predicting these outcomes.[36] Nonetheless, the RCRI poorly predicts all-cause mortality, likely because only 45% of postoperative deaths after noncardiac surgery are due to vascular causes.[9] Similarly, when three risk scores for predicting the need for dialysis after cardiac surgery were instead used to predict less severe grades of acute kidney injury,[43-45] all the risk scores showed diminished prognostic accuracy.[46]

Second, more research is needed to determine how information about expected risks should be communicated to patients, especially in the preoperative setting where this information can influence decisions regarding whether non-operative alternatives to surgery should be considered. In the North American setting, as many as 75% of patients want to participate actively in the decision-making for surgical or invasive procedures.[47] Many anesthesiologists will encounter situations where they quote a very high expected perioperative risk (e.g., 30% chance of dying after surgery) to a patient, yet the same patient views this risk much more optimistically. These differences may be partly explained in part by how physicians express information on risk to patients. Risk might be communicated in several different ways, including qualitatively (e.g., much higher risk than average), graphically, numerically in absolute terms (e.g., absolute risk differences or numbers needed to treat) or numerically in relative terms (e.g., relative risk differences). Importantly, the manner in which risk information is communicated can influence patients' decision-making.[48]

Third, development of new preoperative biomarkers, such as natriuretic peptides or high-sensitivity troponins,[49,50] must be accompanied by research on how best to integrate them with existing risk scores. While an individual biomarker may provide accurate risk information when used in isolation, the key clinical question is whether the new test contributes additional prognostic information beyond that provided by risk scores based on readily available clinical information. Risk reclassification table analyses are integral to making this assessment. *Fourth*, improved methods to predict risk must be accompanied by interventions to improve outcomes in patients within these newly defined risk strata. At present, more accurate identification of low-risk patients facilitates progressing rapidly to surgery without further testing, interventions, or delays, while accurate identification of high-risk patients facilitates allocating more resources to intensive perioperative monitoring for early detection and treatment of adverse events. As indicated above, the key challenge that remains is to identify proven interventions to prevent postoperative complications in high-risk patients. More research is needed to fill the critical gap in the perioperative literature.

Fifth, risk information from preoperative assessment must be better updated based on prognostically important events identified during the intraoperative period. For example, most anesthesiologists will recall being pleasantly surprised by high-risk patients who have an unexpectedly smooth intraoperative course and therefore would now be expected to be at low risk for postoperative

complications. Conversely, many anesthesiologists will also encounter patients who were predicted to be at low risk before surgery but who then experience unexpected intraoperative complications such as major bleeding. Such individuals would be reclassified as being at an elevated risk for postoperative complications due to these intraoperative events. There is an emerging body of research focused on identifying intraoperative events associated with adverse postoperative outcomes, such as hypotension and major bleeding.[51,52] In some cases, this intraoperative information has been integrated into immediate postoperative risk scores, a key example being the Surgical Apgar Score.[51] Future research must evaluate how preoperative risk scores should be integrated with prognostically important intraoperative information, especially to allow for more appropriate limited resources for intensive postoperative monitoring.

Conclusions

Preoperative risk scores can help inform clinical decision-making for patients awaiting surgery. A good risk score should be simple, reliable, and prognostically accurate. When assessing the prognostic accuracy of a risk score, several different analytical methods must be employed to evaluate the relevant domains of discrimination and calibration. Furthermore, all newly developed risk scores, even when initially shown to be prognostically accurate, should still undergo rigorous external validation to assess the stability of this performance when applied to new settings, populations, or outcome events. Future research on preoperative risk scores should determine how preoperative risk estimates can be updated with prognostically important information from the intraoperative period, which novel biomarkers warrant integration into existing risk scores, how risk information should best be communicated to patients, and which interventions can improve outcomes among patients within newly identified risk strata.

Financial support Dr. Wijeysundera is supported in part by a New Investigator Award from the Canadian Institutes of Health Research, and a Merit Award from the Department of Anesthesia at the University of Toronto.

Conflicts of interest None declared.

References

1. *Fleisher LA*, *Fleischmann KE*, *Auerbach AD*, *et al*. 2014 ACC/AHA guideline on perioperative cardiovascular evaluation and management of patients undergoing noncardiac surgery: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014; 130: e278-333.
2. *Lindenauer PK*, *Pekow P*, *Wang K*, *Mamidi DK*, *Gutierrez B*, *Benjamin EM*. Perioperative beta-blocker therapy and mortality after major noncardiac surgery. N Engl J Med 2005; 353: 349-61.
3. *London MJ*, *Hur K*, *Schwartz GG*, *Henderson WG*. Association of perioperative beta-blockade with mortality and cardiovascular morbidity following major noncardiac surgery. JAMA 2013; 309: 1704-13.
4. *Ghaferi AA*, *Birkmeyer JD*, *Dimick JB*. Variation in hospital mortality associated with inpatient surgery. N Engl J Med 2009; 361: 1368-75.
5. *Birkmeyer JD*, *Siewers AE*, *Finlayson EV*, *et al*. Hospital volume and surgical mortality in the United States. N Engl J Med 2002; 346: 1128-37.
6. *Glance LG*, *Kellermann AL*, *Hannan EL*, *et al*. The impact of anesthesiologists on coronary artery bypass graft surgery outcomes. Anesth Analg 2015; 120: 526-33.
7. *Arozullah AM*, *Khuri SF*, *Henderson WG*, *Daley J*, *Participants in the National Veterans Affairs Surgical Quality Improvement Program*. Development and validation of a multifactorial risk index for predicting postoperative pneumonia after major noncardiac surgery. Ann Intern Med 2001; 135: 847-57.
8. *Futier E*, *Constantin JM*, *Paugam-Burtz C*, *et al*. A trial of intraoperative low-tidal-volume ventilation in abdominal surgery. N Engl J Med 2013; 369: 428-37.
9. *Vascular Events In Noncardiac Surgery Patients Cohort Evaluation (VISION) Study Investigators*, *Devereaux PJ*, *Chan MT*, *Alonso-Coello P*, *et al*. Association between postoperative troponin levels and 30-day mortality among patients undergoing noncardiac surgery. JAMA 2012; 307: 2295-304.
10. *Mooney JF*, *Ranasinghe I*, *Chow CK*, *et al*. Preoperative estimates of glomerular filtration rate as predictors of outcome after surgery: a systematic review and meta-analysis. Anesthesiology 2013; 118: 809-24.
11. *Musallam KM*, *Tamim HM*, *Richards T*, *et al*. Preoperative anaemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2011; 378: 1396-407.
12. *Saklad M*. Grading of patients for surgical procedures. Anesthesiology 1941; 2: 281-4.
13. *Laupacis A*, *Sekar N*, *Stiell IG*. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997; 277: 488-94.
14. *Aronson WL*, *McAuliffe MS*, *Miller K*. Variability in the American Society of Anesthesiologists physical status classification scale. AANA J 2003; 71: 265-74.
15. *Cuvillon P*, *Nouvellon E*, *Marret E*, *et al*. American Society of Anesthesiologists' physical status system: a multicentre Francophone study to analyse reasons for classification disagreement. Eur J Anaesthesiol 2011; 28: 742-7.
16. *Haynes SR*, *Lawler PG*. An assessment of the consistency of ASA physical status classification allocation. Anaesthesia 1995; 50: 195-9.
17. *Mak PH*, *Campbell RC*, *Irwin MG*, *American Society of Anesthesiologists*. The ASA physical status classification: inter-observer consistency. Anaesth Intensive Care 2002; 30: 633-40.
18. *Owens WD*, *Felts JA*, *Spitznagel EL Jr*. ASA physical status classifications: a study of consistency of ratings. Anesthesiology 1978; 49: 239-43.
19. *Ranta S*, *Hynynen M*, *Tammisto T*. A survey of the ASA physical status classification: significant variation in allocation among Finnish anaesthesiologists. Acta Anaesthesiol Scand 1997; 41: 629-32.
20. *Sankar A*, *Johnson SR*, *Beattie WS*, *Tait G*, *Wijeysundera DN*. Reliability of the American Society of Anesthesiologists physical status scale in clinical practice. Br J Anaesth 2014; 113: 424-32.

21. *Koo CY*, *Hyder JA*, *Wanderer JP*, *Eikermann M*, *Ramachandran SK*. A meta-analysis of the predictive accuracy of postoperative mortality using the American Society of Anesthesiologists' physical status classification system. World J Surg 2015; 39: 88-103.
22. *Moonesinghe SR*, *Mythen MG*, *Das P*, *Rowan KM*, *Grocott MP*. Risk stratification tools for predicting morbidity and mortality in adult patients undergoing major surgery: qualitative systematic review. Anesthesiology 2013; 119: 959-81.
23. *Charlson ME*, *Pompei P*, *Ales KL*, *MacKenzie CR*. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987; 40: 373-83.
24. *Chertow GM*, *Lazarus JM*, *Christiansen CL*, *et al*. Preoperative renal risk stratification. Circulation 1997; 95: 878-84.
25. *Lapuerta P*, *L'Italien GJ*, *Paul S*, *et al*. Neural network assessment of perioperative cardiac risk in vascular surgery patients. Med Decis Making 1998; 18: 70-5.
26. *Harrell FE Jr*. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis (Springer Series in Statistics). NY: Springer-Verlag Inc.; 2001 .
27. *Kuhn M*, *Johnson K*. Applied Predictive Modeling. NY: Springer-Verlag Inc.; 2013 .
28. *Steyerberg E*. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health). NY: Springer-Verlag Inc.; 2009 .
29. *Moons KG*, *Altman DG*, *Reitsma JB*, *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015; 162: W1-73.
30. *Jaeschke R*, *Guyatt GH*, *Sackett DL*. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994; 271: 703-7.
31. *Leening MJ*, *Vedder MM*, *Witteman JC*, *Pencina MJ*, *Steyerberg EW*. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med 2014; 160: 122-31.
32. *Sheth T*, *Chan M*, *Butler C*, *et al*. Prognostic capabilities of coronary computed tomographic angiography before non-cardiac surgery: prospective cohort study. BMJ 2015; 350: h1907.
33. *Muhlenbruch K*, *Heraclides A*, *Steyerberg EW*, *Joost HG*, *Boeing H*, *Schulze MB*. Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories. Eur J Epidemiol 2013; 28: 25-33.
34. *Ridker PM*, *Buring JE*, *Rifai N*, *Cook NR*. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA 2007; 297: 611-9.
35. *Lee TH*, *Marcantonio ER*, *Mangione CM*, *et al*. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. Circulation 1999; 100: 1043-9.
36. *Ford MK*, *Beattie WS*, *Wijeysundera DN*. Systematic review: Prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. Ann Intern Med 2010; 152: 26-35.
37. *Bilimoria KY, Liu Y, Paruch JL, et al.* Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg 2013; 217: 833-42.e1-3.
38. *Lundstrom LH*, *Vester-Andersen M*, *Moller AM*, *et al*. Poor prognostic value of the modified Mallampati score: a meta-analysis involving 177 088 patients. Br J Anaesth 2011; 107: 659-67.
39. *Karkouti K*, *Rose DK*, *Ferris LE*, *Wigglesworth DF*, *Meisami-Fard T*, *Lee H*. Inter-observer reliability of ten tests used for predicting difficult tracheal intubation. Can J Anaesth 1996; 43: 554-9.
40. *Whiteley MS*, *Prytherch DR*, *Higgins B*, *Weaver PC*, *Prout WG*. An evaluation of the POSSUM surgical scoring system. Br J Surg 1996; 83: 812-5.
41. *Prytherch DR*, *Whiteley MS*, *Higgins B*, *Weaver PC*, *Prout WG*, *Powell SJ*. POSSUM and Portsmouth POSSUM for predicting mortality. Physiological and operative severity score for the enUmeration of mortality and morbidity. Br J Surg 1998; 85: 1217-20.
42. *Sutton R*, *Bann S*, *Brooks M*, *Sarin S*. The surgical risk scale as an improved tool for risk-adjusted analysis in comparative surgical audit. Br J Surg 2002; 89: 763-8.
43. *Mehta RH*, *Grab JD*, *O'Brien SM*, *et al*. Bedside tool for predicting the risk of postoperative dialysis in patients undergoing cardiac surgery. Circulation 2006; 114: 2208-16.
44. *Thakar CV*, *Arrigain S*, *Worley S*, *Yared JP*, *Paganini EP*. A clinical score to predict acute renal failure after cardiac surgery. J Am Soc Nephrol 2005; 16: 162-8.
45. *Wijeysundera DN*, *Karkouti K*, *Dupuis JY*, *et al*. Derivation and validation of a simplified predictive index for renal replacement therapy after cardiac surgery. JAMA 2007; 297: 1801-9.
46. *Englberger L*, *Suri RM*, *Li Z*, *et al*. Validation of clinical scores predicting severe acute kidney injury after cardiac surgery. Am J Kidney Dis 2010; 56: 623-31.
47. *Mazur DJ*, *Hickam DH*. Patients' preferences for risk disclosure and role in decision making for invasive medical procedures. J Gen Intern Med 1997; 12: 114-7.
48. *Fagerlin A*, *Zikmund-Fisher BJ*, *Ubel PA*. Helping patients decide: ten steps to better risk communication. J Natl Cancer Inst 2011; 103: 1436-43.
49. *Rodseth RN*, *Biccard BM*, *Le Manach Y*, *et al*. The prognostic value of pre-operative and post-operative B-type natriuretic peptides in patients undergoing noncardiac surgery: B-type natriuretic peptide and N-terminal fragment of pro-B-type natriuretic peptide: a systematic review and individual patient data meta-analysis. J Am Coll Cardiol 2014; 63: 170-80.
50. *Weber M*, *Luchner A*, *Seeberger M*, *et al*. Incremental value of high-sensitive troponin T in addition to the revised cardiac index for peri-operative risk stratification in non-cardiac surgery. Eur Heart J 2013; 34: 853-62.
51. *Gawande AA*, *Kwaan MR*, *Regenbogen SE*, *Lipsitz SA*, *Zinner MJ*. An Apgar score for surgery. J Am Coll Surg 2007; 204: 201-8.
52. *Walsh M*, *Devereaux PJ*, *Garg AX*, *et al*. Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: toward an empirical definition of hypotension. Anesthesiology 2013; 119: 507-15.
53. *Copeland GP*, *Jones D*, *Walters M*. POSSUM: A scoring system for surgical audit. Br J Surg 1991; 78: 355-60.
54. *Donati A*, *Ruzzi M*, *Adrario E*, *et al*. A new and feasible model for predicting operative risk. Br J Anaesth 2004; 93: 393-9.
55. *Mallampati SR*, *Gatt SP*, *Gugino LD*, *et al*. A clinical sign to predict difficult tracheal intubation: a prospective study. Can Anaesth Soc J 1985; 32: 429-34.
56. *Roques F*, *Nashef SA*, *Michel P*, *et al*. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg 1999; 15: 816-22 **discussion 822-3**.
57. *Dupuis JY*, *Wang F*, *Nathan H*, *Lam M*, *Grimes S*, *Bourke M*. The cardiac anesthesia risk evaluation score: a clinically useful predictor of mortality and morbidity after cardiac surgery. Anesthesiology 2001; 94: 194-204.
58. *Marcantonio ER*, *Goldman L*, *Mangione CM*, *et al*. A clinical prediction rule for delirium after elective noncardiac surgery. JAMA 1994; 271: 134-9.

59. *Knaus WA*, *Draper EA*, *Wagner DP*, *Zimmerman JE*. APACHE II: a severity of disease classification system. Crit Care Med 1985; 13: 818-29.
60. *Malinchoc M*, *Kamath PS*, *Gordon FD*, *Peine CJ*, *Rank J*, *ter Borg PC*. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. Hepatology 2000; 31: 864-71.