European Journal of Cardio-Thoracic Surgery 54 (2018) 203–208
doi:10.1093/ejcts/ezy180 Advance Access publication 7 May 2018

REVIEW

Cite this article as: Grant SW, Collins GS, Nashef SAM. Statistical Primer: developing and validating a risk prediction model. Eur J Cardiothorac Surg 2018;54:203–8.

STATISTICAL PRIMER

Statistical Primer: developing and validating a risk prediction model†

Stuart W. Grant[a,*], Gary S. Collins[b] and Samer A.M. Nashef[c]

[a] Department of Academic Surgery, University of Manchester, Manchester, UK
[b] Centre for Statistics in Medicine, University of Oxford, Oxford, UK
[c] Department of Cardiothoracic Surgery, Papworth Hospital, Cambridge, UK

* Corresponding author. Department of Academic Surgery, Wythenshawe Hospital, University of Manchester, Manchester M23 9LT, UK. Tel: +44-0161-2915853; e-mail: stuart.grant@manchester.ac.uk (S.W. Grant).

Received 1 December 2017; received in revised form 30 March 2018; accepted 2 April 2018

Summary

A risk prediction model is a mathematical equation that uses patient risk factor data to estimate the probability of a patient experiencing a healthcare outcome. Risk prediction models are widely studied in the cardiothoracic surgical literature with most developed using logistic regression. For a risk prediction model to be useful, it must have adequate discrimination, calibration, face validity and clinical usefulness. A basic understanding of the advantages and potential limitations of risk prediction models is vital before applying them in clinical practice. This article provides a brief overview for the clinician on the various issues to be considered when developing or validating a risk prediction model. An example of how to develop a simple model is also included.

Keywords: Risk prediction • Calibration • EuroSCORE • EuroSCORE II • Receiver operating characteristic • Risk assessment • Surgical mortality

INTRODUCTION

A risk prediction model is a mathematical equation that uses patient risk factor data to estimate the probability of a patient experiencing a healthcare outcome. Risk prediction models are used throughout medical practice for a variety of purposes such as predicting development of a disease, predicting response to treatment or predicting patient prognosis. In surgery, they are commonly used to predict the risk of adverse outcomes after intervention. Surgical risk prediction models can be used to facilitate clinical decision making, define thresholds for intervention or for risk-adjusting outcome data for benchmarking purposes. The most well-known and widely studied models in cardiothoracic surgery are the EuroSCORE models and the procedure-specific Society of Thoracic Surgery (STS) models [1–3].

There are many different statistical techniques that can be used to develop a risk prediction model including but not limited to logistic regression, linear regression, Cox regression and machine learning. The outcome can also be either binary or continuous. Most risk prediction models in the cardiothoracic literature are developed using multivariable logistic regression to predict a binary outcome. As a result, although many of the core principles discussed throughout this article apply to all risk prediction models, the focus is on risk prediction models based on logistic regression for a binary outcome. The key issues to consider when developing and validating a risk prediction model are summarized in Table 1 and described in more detail below.

†Presented at the 31st Annual Meeting of the European Association for Cardio-Thoracic Surgery, Vienna, Austria, 7–11 October 2017.

METHODOLOGY

Model objective

Before starting to develop a risk prediction model, it is important to consider whether a new model is needed. The literature should be reviewed to identify, evaluate and potentially consider updating any existing models. Once a new risk prediction model is deemed necessary, developing it is a balancing act between clinical usefulness, statistical performance and functionality. It is important that the objective of the model is clearly defined, so that all these aspects can be balanced appropriately to meet the objective. For example, a model to risk-adjust surgical outcome data across a range of procedures with the most accurate possible statistical performance for benchmarking purposes should be quite different from a model designed to allow patients to be able to estimate their risk of developing a complication after a specific procedure. The first model should cover multiple procedures and may be statistically quite complex containing multiple predictors. The second model should be procedure specific and only certain predictors that would be accessible to patients should be included. It is important, therefore, that the model objective is carefully considered during each aspect of model development.

Data considerations

Once the need to develop a model has been determined and the objective established, it is important to ensure that there are

© The Author(s) 2018. Published by Oxford University Press on behalf of the European Association for Cardio-Thoracic Surgery. All rights reserved.

Downloaded from https://academic.oup.com/ejcts/article/54/2/203/4993384 by University of Oklahoma user on 13 October 2021

Table 1: Issues to consider when developing a risk prediction model

Issues	Comments
Existing models	Consider whether a new model is needed. Identify, evaluate and potentially update any existing models
Candidate predictors	Only consider predictors that have a plausible relationship with the outcome. Use subject (clinical) knowledge and systematic reviews to identify candidate predictors
Sample size	Ensure the ratio of number of outcome events to the number of parameters that could potentially be estimated is at least 10
Continuous predictors	Avoid categorizing continuous predictors and consider non-linear relationships with the outcome (e.g. using restricted cubic splines or fractional polynomials)
Missing data	Avoid omitting individuals with incomplete data. Consider using multiple imputation
Overfitting	Consider shrinkage or penalization methods (e.g. lasso) to limit overfitting
Internal validation	Avoid randomly splitting data into development and validation and use cross-validation or bootstrapping. Ensure all variable selection is captured in the internal validation
External validation	Evaluating model performance on other data sets is important to judge generalizability and transportability
Model performance	Evaluate both discrimination and calibration
Model impact	Assessed through net benefit methods or impact studies
Reporting	Ensure all key details are transparently reported, following the TRIPOD reporting guidelines

sufficient data available to develop the model and that the data are of good enough quality. Although a considerable number of risk prediction models are developed using existing data, prospectively designed studies are the best approach to ensure that both data quantity and quality are adequate. With regard to study size, current recommendations are that at least 10 outcomes are required for the investigation of one candidate predictor for inclusion in a multivariable model [4]. More precisely, this recommendation applies to the number of regression coefficients for all candidate predictors that require estimating rather than just the number of candidate predictors. As a result, a candidate predictor with four categories requires three regression coefficients to be estimated, meaning at least 30 outcomes would be required for the investigation of this candidate predictor. Therefore, if the model is being developed for an infrequently occurring outcome such as mortality after cardiac surgery, then a large overall data set is required for model development. If the outcome occurs more frequently, then smaller data sets may be used. Modern machine learning approaches may need over 10 times the number of outcomes when compared with traditional methods [5].

Missing data, particularly for clinically important predictors, should be kept to a minimum. If missing data are present, many strategies are available to deal with missing data, but each has limitations. A complete case analysis can substantially reduce the data available for model development and lead to inaccurate estimates of specific predictors or overall model performance. Multiple imputation, which maintains the size of the data set available for model development, is the preferred approach but relies on the assumption that the data are missing at random. If a predictor is missing in a sizable proportion of the data, then it is questionable whether inclusion of the predictor in the model is worthwhile [6]. Discarding predictors based on the amount of missing data is largely a subjective decision; however, if a predictor has a high proportion of missing data in the development data set, then the influence of the predictor and the model performance may be inaccurately estimated [7].

Data should be representative of the population in which the risk prediction model is intended to be used. For example, if the objective is to develop a model to be used across different geographical areas, then it is important that these geographical areas are represented in the development data. In addition, it is important that the data are as contemporaneous as possible. Clinical practice changes over time and, therefore, models that have been developed with historical data or data collected over an extended period of time may demonstrate inadequate performance [8].

Predictors

Predictors for inclusion in the risk prediction model may be identified by expert opinion or through a review of the literature. Predictors for inclusion in the multivariable model are often identified by assessing their univariable association with the outcome. However, excluding potentially useful risk factors merely because they are not significantly associated with the outcome on univariable analysis is not recommended [6]. The inclusion of strong predictors is essential for the development of useful risk prediction models. The strength of a predictor is related to not only the association between the predictor and the outcome but also the distribution of the predictor in the development data. A predictor that is strongly associated with the outcome but only occurs in a small number of patients would not be as strong as a risk factor with a slightly smaller association with the outcome but that is present in half of the patients. This is the reason why some rare predictors that are strongly associated with outcomes after cardiac surgery, such as hepatic dysfunction [9], are not incorporated in the EuroSCORE models.

For surgical risk prediction models, predictors commonly include patient demographics (e.g. age and gender), comorbidities (e.g. diabetes and respiratory disease), previous relevant medical history (e.g. previous surgery), functional or symptom information (e.g. angina or dyspnoea grade) laboratory investigations (e.g. creatinine), imaging (e.g. left ventricular (LV) function and left main disease) and procedure urgency or complexity. For a predictor to be potentially useful, it should be objectively measured, easily available, clearly defined and have minimal measurement error. Predictors that are highly correlated are unlikely to contribute significant independent information to the multivariable model and one or the other should generally be excluded [10]. Some predictors may have an enhanced effect on the outcome when present in combination. Such a phenomenon is called an interaction and may be included in the model via an interaction term if it is significant. Although significant interaction terms may be identified, inclusion of them in the model does not necessarily improve model performance [11].

The number of groups for a categorical predictor can be collapsed if there are no outcomes in one of the categories.

Downloaded from https://academic.oup.com/ejcts/article/54/2/203/4993384 by University of Oklahoma user on 13 October 2021

However dichotomization of a continuous variable should be avoided as this can reduce the power by approximately the same amount as discarding one-third of the data [12]. If a decision to dichotomize or group a continuous predictor is made, then this should be done using predefined thresholds. The relationship between any continuous predictors and the outcome should be assessed. If a continuous predictor does not demonstrate a linear relationship with the outcome, modelling of the predictor to capture the non-linear relationship should be performed (e.g. using fractional polynomials or restricted cubic splines).

Outcome

As with predictors, for an outcome to be useful, it should be easily available, clearly defined and have minimal measurement error. It should also be of importance to the patient, clinician and healthcare provider. Perioperative mortality is a commonly used outcome, and although it is clearly important and has zero measurement error, there may be issues with data availability and definition. Perioperative mortality could be defined as death in the hospital irrespective of timing or could be death within a certain time from the operation regardless of location or a combination of the two. There may be potentially important differences in model's performance if a model was developed for one perioperative mortality outcome but validated on another. While a perioperative mortality outcome based on time such as 30-day or 90-day mortality has the benefit of providing a consistent follow-up period for all patients, only half of hospitals involved in the EuroSCORE II project had access to this data. As a result, in-hospital mortality was the preferred outcome for the EuroSCORE II model [1].

Other outcomes may potentially be more relevant for risk prediction models focusing on specific procedures with a very low incidence of perioperative mortality. Such outcomes after cardiac surgery could include postoperative stroke, renal injury, wound infection, the need for transfusion, reoperation for bleeding and postoperative length of stay. While these short-term outcomes are usually easily available and important, there are often issues regarding clear outcome definitions. While more long-term outcomes after surgical intervention can be highly important markers of quality, developing risk prediction models based on these outcomes are often limited by data availability.

Model development

Most risk prediction models in the cardiothoracic literature are developed using logistic regression. Logistic regression allows both categorical and continuous predictors to be included in a model to predict a binary outcome with the predictions from the model bounded by 0 and 1. Numerous statistical packages capable of performing logistic regression analyses, such as SPSS, R and SAS, and an example of a basic model development is shown below.

Historically, logistic models were often converted into simple additive scores to make them easier to use by assigning weights to the predictors based on the log odds ratios obtained from the model. Given improvements in access to software designed to calculate logistic scores, this approach is now generally not required. It is also not recommended because there is usually no link between the additive score and the actual predicted risk, it does not allow the incorporation of continuous predictors and discriminatory ability is usually compromised. Other approaches to model development include machine learning approaches such as neural networks. However, these approaches require substantially larger sample sizes, are prone to overfitting and are more complex to interpret. Empirical studies have compared machine learning approaches with more traditional regression with little difference between the two approaches [13].

For logistic regression, the two main strategies for development of the final model are full model and stepwise selection. In the full model approach, all predictors are included in the model irrespective of their association with the outcome or influence on model performance. In the stepwise selection approach, predictors are removed or added to the model based on a sequence of hypothesis tests. Backward model selection where all predictors are included at first and predictors are subsequently removed is generally preferred to forward model selection, whereby the model is built up by adding predictors in starting with the strongest predictor. Although stepwise selection may be useful, a potential limitation of model selection strategies is that it can lead to overfitting of the model.

Overfitting means that the model is too specific to the development data and may not be generalizable outside the development cohort because random variation present in the development data set is captured along with any clinical associations between the outcomes and the predictors. Models that are overfitted will perform poorly on external validation. The full model approach may reduce overfitting, but it is often impractical in data sets with large numbers of candidate predictors. Penalized methods such as the lasso or elastic net can reduce overfitting during the model building process by shrinking the regression coefficients [14].

Example

The objective is to develop a risk prediction model for in-hospital mortality to be applied to patients undergoing all types of cardiac surgery based on only age, gender, LV function and operative urgency. Data are available for 14 017 procedures. The patient characteristics data are shown in Table 2. The in-hospital mortality rate for the cohort is 2.4%. Logistic regression with no model selection strategy is applied to the data. The output of the logistic regression analysis is shown in Table 3. To calculate the risk of in-hospital mortality for an individual patient, the following calculations need to be performed with the coefficient values multiplied by 1 if the risk factor is present and by 0 if absent.

$$\begin{aligned}&\text{linear predictor (LP)} = -9.605 + (0.073 * \text{age in years})\\&+ (0.463 * \text{female})\\&+ [0.585 * \text{moderate LV function})\\&+ (1.294 * \text{poor LV function}) + (0.559 * \text{urgent})\\&+ (2.528 * \text{emergency})\end{aligned}$$

$$\begin{aligned}&\text{Predicted risk of in-hospital mortality}\\&= 1/(1 + \exp(-\text{LP}))\end{aligned}$$

Statistical performance

Statistical performance of a risk prediction model is assessed across two main characteristics: discrimination and calibration.

Downloaded from https://academic.oup.com/ejcts/article/54/2/203/4993384 by University of Oklahoma user on 13 October 2021

Discrimination assesses how well the model differentiates between those patients who experience the outcome and those who do not. It is commonly measured by the area under the receiver operating characteristic curve. For a binary outcome, this is equivalent to the concordance statistic (*c*-statistic). The receiver operating characteristic curve is a plot of sensitivity (true-positive rate) against 1 - specificity (false-positive rate) for consecutive cut-offs for the predicted risk [15]. An area under the receiver operating characteristic curve of 0.50 indicates that the model is no better than a random guess at assigning higher scores to those patients who experience the outcome than those who do not experience the outcome. Although arbitrary, values ≥0.70 are generally considered to be useful, with values ≥0.80 considered to be excellent. If a model does not accurately discriminate, then it is not useful as a risk prediction model.

Table 2: Patient characteristics for an example of a model development cohort (n=14017)

Patient characteristics	*n*/mean	%/SD
Age[a] (years)	66.7	10.8
Female	3819	27.2
LV function		
Good	10 958	78.2
Moderate	2401	17.1
Poor	658	4.7
Urgency		
Elective	10 477	74.7
Urgent	3214	22.9
Emergency	326	2.3
Logistic EuroSCORE	6.1	7.8

[a]Continuous predictors.

LV: left ventricular; SD: standard deviation.

Table 3: The logistic regression model for in-hospital mortality after cardiac surgery based on an example of a development data set

Patient characteristics	Coefficient	Odds ratio (95% CI)	*P*-value
Age[a] (years)	0.073	1.076 (1.062–1.090)	<0.001
Female	0.463	1.589 (1.260–2.005)	<0.001
LV function			
Good			<0.001
Moderate	0.585	1.795 (1.378–2.339)	<0.001
Poor	1.294	3.646 (2.579–5.514)	<0.001
Urgency			
Elective			<0.001
Urgent	0.559	1.749 (1.358–2.252)	<0.001
Emergency	2.528	12.528 (8.822–17.793)	<0.001
Intercept	-9.065		<0.001

[a]Continuous predictor.

CI: confidence interval; LV: left ventricular.

Calibration is an assessment of how closely the predictions of the model match the observed outcomes in the data. Calibration can be assessed in several ways. Unlike for discrimination, if a model is poorly calibrated, there are methods that can be used to recalibrate the model appropriately [16]. The most simple way is to calculate the observed to expected (*O*:*E*) ratio by dividing the mean observed and predicted outcome rates. If a risk prediction model is perfectly calibrated, the *O*:*E* ratio would be 1. *O*:*E* ratios above or below one are indicative of under-prediction and over-prediction, respectively. The *O*:*E* ratio on its own does not provide adequate information regarding model calibration as over-prediction in one subgroup could be combined with under-prediction in another to give an overall value close to one.

Model calibration can be nicely assessed graphically using a calibration plot as shown in Fig. 1. In a calibration plot, the mean predicted probability of outcome is plotted against the observed proportion of outcomes for groups (normally 10) of the cohort. The line of equality which represents perfect calibration should be overlaid, approximate 95% confidence intervals for the observed mortality can be displayed as error bars and smoothing techniques can be used to depict the association between the observed and predicted outcomes. Calibration can also be assessed by fitting a logistic regression model with the outcome variable set as the observed outcomes and the independent variable set as the log-odds transformed model predictions. If the model is perfectly calibrated, then the model intercept and slope would equal 0 and 1, respectively. The Hosmer–Lemeshow test is often used to assess model calibration and involves splitting the cohort, often into 10 equally sized groups, with contributing χ^2 statistics from each group then summed to give an overall *P*-value [17]. However, the test is influenced by the sample size, the number of groups and provides no information on the direction or magnitude of miscalibration.

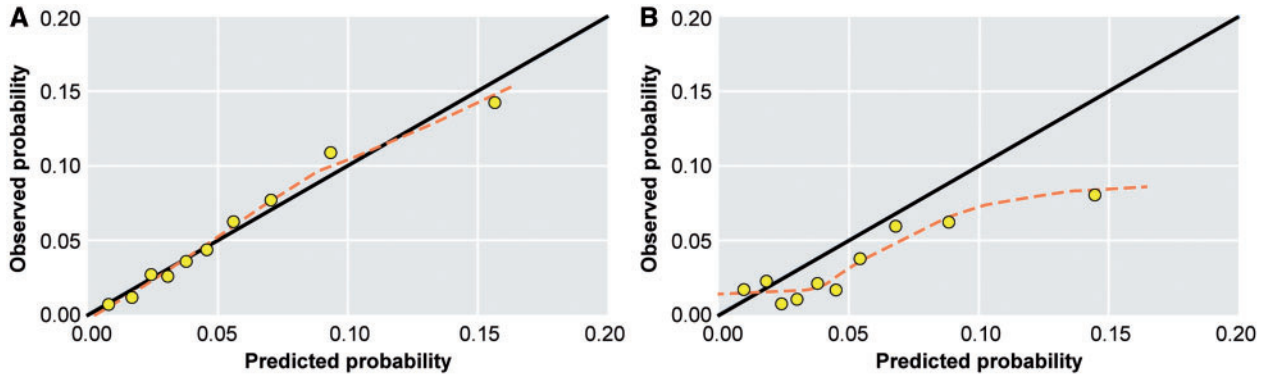

Figure 1: Two example calibration plots showing the mean predicted probability of outcomes plotted against the observed proportion of outcomes for 10 equally sized groups (yellow dots). The black line is at 45° and represents a line of equality and perfect calibration. The dashed red line is a smoothed locally weighted scatterplot smoothing (LOWESS) regression line. (**A**) A well-calibrated model with almost all points close to the line of equality. (**B**) A poorly calibrated model with over-prediction of risk in the majority of groups.

Downloaded from https://academic.oup.com/ejcts/article/54/2/203/4993384 by University of Oklahoma user on 13 October 2021

Model validation

Risk prediction models should be both internally and externally validated before they are adopted in clinical practice [18]. Internal model validation is the process of assessing optimism and quantifying statistical performance of the model using the data on which the model was developed. The performance of a risk prediction model in the data sample from which it was developed is likely to be over optimistic. The preferred approach for internal validation is to use bootstrapping or *k*-fold cross-validation. It is important that the same model building steps used to develop the model are replayed in the bootstrapping or cross-validation. Merely evaluating the final model in different bootstrap samples or cross-validation folds will lead to biased estimates of the optimism. An alternative internal validation approach, whereby the data are randomly split into development and validation data, is inefficient. For small to moderately sized data, it reduces the sample size for model development, therefore increasing the chances of overfitting, and leaves too few data to evaluate the model. For large data sets, randomly splitting data merely creates two comparable data sets and is, therefore, not a strong test of model performance.

External validation (where the statistical performance of a risk prediction model is assessed in a new but similar cohort of patients) is the strongest test of a model. External validation can be performed in different geographical areas, for different time periods or even potentially for different outcomes. A good model should retain good statistical performance across a range of settings and for comparable outcomes such as in-hospital or 30-day mortality. If a model demonstrates poor discrimination on external validation, then it is likely that a new model is required; however, if a model demonstrates poor calibration, it can potentially be updated or recalibrated [16]. If a model consistently demonstrates poor calibration, then it is likely that a new model is required.

Face validity, clinical usefulness and application

Face validity and clinical usefulness should be considered alongside statistical performance for all risk prediction models designed to be applied in clinical practice. Although there is no way to formally assess a model's face validity, there are a number of features that could bring the face validity of a model into question. For example, face validity may be questioned if predictors or interactions are included in the model that would not be expected to be associated with the outcome based on previous research. The face validity of a model may also be questioned if key predictors are not included because they were simply not available in the development data. If the definitions of the predictors or outcomes are unclear or ambiguous, then this will raise concerns about the face validity and limit the application of the model.

When thinking about the clinical usefulness of a risk prediction model, both the applicability of the model to contemporary clinical practice and the additional benefit of using the model above in current practice should be considered. If a model is based on data that do not represent contemporary practice or the model is so specific to a particular situation that it is unlikely to be applicable more generally, then this will limit the clinical usefulness of the model. Accurate performance of the model in key clinical subgroups is also important. If feasible, assessment of model performance in key clinical subgroups could be performed during model development or validation studies. One such example is the assessment of the EuroSCORE models in emergency surgery [19]. The ultimate test of a risk prediction models clinical usefulness is through an impact study that assess the impact of the risk prediction model on clinical practice. Impact studies should ideally be randomized and can either be assistive which is where the model predicted probabilities are provided to the clinician to assist in decision making or decisive where the clinical decision is explicitly decided by the model [20].

Although impact studies are often difficult to conduct, it is possible to assess how much a risk prediction model adds to current standards or existing prediction models using net benefit and decision curve analysis [21]. Net benefit decision curve analysis allows the implications of basing decisions to operate on the predictions generated from the risk prediction model across a range of predicted risks to be compared using a common scale. An example of a decision curve analysis comparing the original logistic EuroSCORE with a recalibrated logistic EuroSCORE and

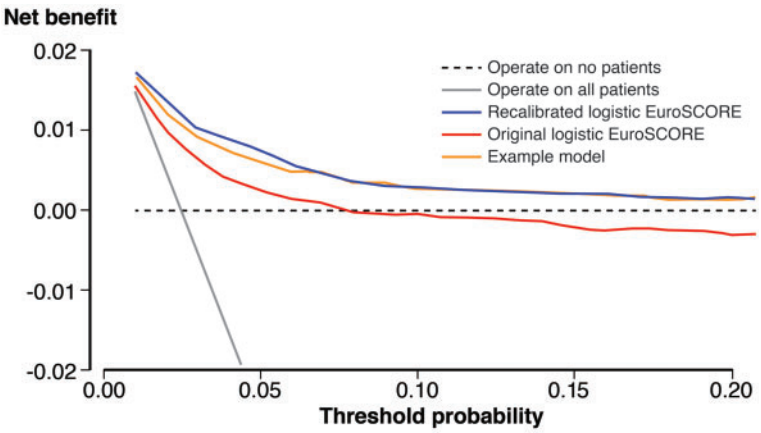

Figure 2: Decision curves showing the clinical usefulness of the original logistic EuroSCORE, a recalibrated logistic EuroSCORE and the model developed in this article for predicting in-hospital mortality. The range of threshold probabilities is set to a maximum of 20% on the *x*-axis with the net benefit displayed on the *y*-axis. The grey line represents the net benefit of performing surgery for all patients, and the dark black line represents performing surgery on no patients. The black dashed line, green dashed line and red dashed line represent the net benefit of applying surgery to patients according to the recalibrated logistic EuroSCORE, the example model and the original logistic EuroSCORE, respectively. The basic interpretation is that the model with the highest net benefit at a particular threshold probability has the highest clinical value.

Downloaded from https://academic.oup.com/ejcts/article/54/2/203/4993384 by University of Oklahoma user on 13 October 2021

the simple example model developed in this article is shown in Fig. 2. As would be expected basing the decision to operate on the recalibrated logistic EuroSCORE demonstrates a higher net benefit than both the simple example model and the original logistic EuroSCORE. Basing the decision to operate on the original logistic EuroSCORE in this cohort would result in net harm for predicted risks of approximately 8% and above which again is as expected because the original logistic EuroSCORE is poorly calibrated for contemporary cardiac surgery [8].

When applying risk prediction models in clinical practice, it should be remembered that the risk prediction model gives a probability of the patient experiencing the outcome based on a group of patients with similar characteristics. Even in large data sets, if multiple predictors are included in the model, there will be patients with specific sets of predictors that are not encountered in the development data. It should also be remembered that with logistic regression, the model prediction which ranges between 0 and 1 will always be wrong as the patient will either experience the outcome or will not experience the outcome.

REPORTING

When developing a risk model, it is important that the full prediction model with all regression coefficients and the model intercept is published to allow predictions for individuals to be calculated. Historically, the quality of reporting for risk prediction model development and validation studies has been poor. As a result, the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) recommendations were developed and published in 2015 [22, 23]. The TRIPOD guidelines are a checklist of 22 items deemed essential for transparent reporting of a prediction model study and are designed to improve the quality of risk prediction model research. They are available at https://www.tripod-statement.org/ (last accessed 21 April 2018).

CONCLUSIONS

Risk prediction models have many potential applications in surgical patients. They can be used to facilitate clinical decision making, define thresholds for intervention and for risk-adjusting outcome data. There are numerous factors to consider when developing a risk prediction model including the objective of the model, data quality, predictors available, statistical methodology and the outcome. Although other methods are available, the most common methodology for developing risk prediction models in the cardiothoracic literature is logistic regression. When validating a risk prediction model, discrimination, calibration, face validity and clinical usefulness should all be considered. When undertaking studies on risk prediction models, the TRIPOD guidelines should be followed to ensure that the usefulness of the prediction models studied can be adequately assessed.

Conflict of interest: none declared.

REFERENCES

[1] Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR *et al*. EuroSCORE II. Eur J Cardiothorac Surg 2012;41:734–45.
[2] Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999;16:9–13.
[3] Jin R, Furnary AP, Fine SC, Blackstone EH, Grunkemeier GL. Using Society of Thoracic Surgeons risk models for risk-adjusting cardiac surgery results. Ann Thorac Surg 2010;89:677–82.
[4] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
[5] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14:137.
[6] Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009;338:b604.
[7] Gorelick MH. Bias arising from missing data in predictive models. J Clin Epidemiol 2006;59:1115–23.
[8] Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K *et al*. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. Eur J Cardiothorac Surg 2013;43:1146–52.
[9] Dimarakis I, Grant S, Corless R, Velissaris T, Prince M, Bridgewater B *et al*. Impact of hepatic cirrhosis on outcome in adult cardiac surgery. Thorac Cardiovasc Surg 2015;63:58–66.
[10] Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer, 2001.
[11] Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation and Updating. New York: Springer, 2008.
[12] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127–41.
[13] Nilsson J, Ohlsson M, Thulin L, Höglund P, Nashef SAM, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. J Thorac Cardiovasc Surg 2006;132:12–19.
[14] Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M *et al*. How to develop a more accurate risk prediction model when there are few events. BMJ 2015;351:h3868.
[15] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
[16] Su T-L, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. Stat Methods Med Res 2018;27:185–97.
[17] Hosmer D, Lemeshow S. Applied Logistic Regression. New York: John Wiley, 1989.
[18] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.
[19] Grant SW, Hickey GL, Dimarakis I, Cooper G, Jenkins DP, Uppal R *et al*. Performance of the EuroSCORE Models in Emergency Cardiac Surgery. Circ Cardiovasc Qual Outcomes 2013;6:178–85.
[20] Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.
[21] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6.
[22] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63.
[23] Moons KG, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW *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.

Downloaded from https://academic.oup.com/ejcts/article/54/2/203/4993384 by University of Oklahoma user on 13 October 2021