Curr Anesthesiol Rep (2016) 6:267–275
DOI 10.1007/s40140-016-0171-8

Risk Prediction Models in Perioperative Medicine: Methodological Considerations

Gary S. Collins[1] · Jie Ma[1] · Stephen Gerry[1] · Eric Ohuma[1] · Lang'O Odondi[1] · Marialena Trivella[1] · Jennifer De Beyer[1] · Maria D. L. A. Vazquez-Montes[1,2]

Published online: 5 July 2016

Abstract

Purpose of Review Risk prediction models hold enormous potential for assessing surgical risk in a standardized, objective manner. Despite the vast number of risk prediction models developed, they have not lived up to their potential. The aim of this paper is to provide an overview of the methodological issues that should be considered when developing and validating a risk prediction model to ensure a useful, accurate model.

Recent Findings Systematic reviews examining the methodological and reporting quality of these models have found widespread deficiencies that limit their usefulness.

Summary Risk prediction modelling is a growing field that is gaining huge interest in the era of personalized medicine. Although there are no shortcuts and many challenges are faced when developing and validating accurate, useful prediction models, these challenges are surmountable, if the abundant methodological and practical guidance available is used correctly and efficiently.

Keywords Risk prediction · Discrimination · Calibration · Multivariable · Statistical methods

Introduction

Risk prediction models hold enormous potential for assessing surgical risk in a standardized, objective manner [1]. They can be used to guide clinical decision making and perioperative management, enable informed consent, stratify risk for inclusion into randomized controlled trials, and audit, monitor, assess, and compare surgical outcomes in different healthcare providers [2, 3]. Regardless of the reasons for using a particular risk prediction model, it is important that it is appropriately developed and validated [2, 4, 5].

The last 10–15 years have seen an explosion in risk scores predicting surgical outcomes, such as mortality [6], complications [7], morbidities [8], and bleeding [9]. More widely, risk prediction models have proliferated in the medical literature, resulting in many competing models for the same outcome or target population. For example, there are nearly 800 models for patients with cardiovascular disease [10], over 360 models for predicting incident

This article is part of the Topical collection on *Research Methods and Statistical Analyses*.

✉ Gary S. Collins
gary.collins@csm.ox.ac.uk

Jie Ma
jie.ma@ndorms.ox.ac.uk

Stephen Gerry
stephen.gerry@csm.ox.ac.uk

Eric Ohuma
eric.ohuma@csm.ox.ac.uk

Lang'O Odondi
lango.odondi@oncology.ox.ac.uk

Marialena Trivella
marialena.trivella@csm.ox.ac.uk

Jennifer De Beyer
jennifer.de-beyer@csm.ox.ac.uk

Maria D. L. A. Vazquez-Montes
maria.vazquezmontes@phc.ox.ac.uk

[1] Nuffield Department of Orthopaedics, Centre for Statistics in Medicine, Rheumatology and Musculoskeletal Sciences, University of Oxford, Windmill Road, Oxford OX3 7LD, UK

[2] Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

cardiovascular disease [11], 263 models in obstetrics (with 69 predicting the risk of preeclampsia) [12], over 100 models for patients with prostate cancer [13], and 20 models for predicting prolonged intensive care stay following cardiac surgery [14].

Despite the vast number of risk prediction models developed, they have not lived up to their potential. Systematic reviews examining the methodological and reporting quality of these models have found widespread deficiencies that limit their usefulness [12, 14–15, 16•, 17–21]. The aim of this paper is to provide an overview of the methodological issues that should be considered when developing and validating a risk prediction model to ensure a useful, accurate model. We use the EuroSCORE model as our case study [1]. The development of the original EuroSCORE was described in two separate articles, one identifying the risk factors [22] and one to construct the model [23].

Assessing the Need for a New Risk Prediction Model

Before deciding to build a new risk prediction model, it is useful to check whether there are any existing models that predict similar outcomes in your target population and clinical setting to avoid duplication of effort. If such models do already exist, then you should first evaluate and compare their predictive performance on your data [24]. If they show promising performance, then recalibration or updating may produce the model that you need [25, 26]. A new model should only be developed from scratch if a similar model does not exist or if similar models cannot be recalibrated to meet the needs of your particular target population and clinical setting. The CHARMS checklist provides guidance on how to conduct systematic reviews of risk prediction models, including how to search, what information to extract, and how to assess study quality and risk of bias [27•].

Overview of Steps in Developing and Validating a Risk Prediction Model

If you have access to an existing dataset and do not have to prospectively collect data, then developing a risk prediction model is easy. You can load the data into your statistical software package, click a few buttons, and churn out yet another new model [28]. However, just because you can easily develop a model, it does not mean you should. The resulting model may add an extra line to your list of publications, but it is (hopefully) highly unlikely that an ill-thought-out model will ever be used on an actual patient.

As with any research study, there should be a clear rationale for why a new risk prediction model is needed. A detailed protocol describing every step needed to develop and validate the model should be written and, if possible, published (for example, in diagnprognres.biomedcentral.com) [29•]. The abundant methodological and practical guidance now available to investigators wishing to develop or validate a risk prediction model leaves little excuse for producing unusable models [4, 5, 30–38, 39••, 40, 41•]. Table 1 gives a brief overview of the main issues which are discussed in more detail throughout the article.

Design

An appropriate study design is the key for developing or validating risk prediction models. The preferred design for both development and validation studies is a prospective longitudinal cohort study. This design gives the investigator full control to ensure all relevant predictors and outcomes are measured and collected, thereby minimizing missing values and loss to follow-up. However, risk prediction models often have to be developed and validated using existing data collected for a different purpose. Although using existing data is cost efficient and convenient, these datasets have clear problems. They are often small, include too few outcome events, have missing values, do not include important predictors, or use inaccurate methods for measuring important predictors. Data from randomized clinical trials can be used, but trials' strict eligibility criteria can often limit their generalizability, and the issue of how to handle treatment assignment needs to be addressed [42, 43]. Case–control studies are generally not appropriate for developing prediction models as the correct baseline risk or hazard cannot be estimated from the data [44] unless a nested case–control or case-cohort design is used [44, 45].

EuroSCORE was developed using a prospective cohort study, involving 132 centres from 8 European countries [22, 23]. All patients ($n = 20{,}014$) undergoing cardiac surgery between September and December 1995 were included, with 984 (with approximately 5 % of the cohort omitted after error checking and quality control), leaving 19,030 for analysis [22]. Using a prospective cohort design enabled efficient collection of 97 preoperative and operative risk factors that were deemed credible, objective, and reliable.

Sample Size

Sample size recommendations for studies developing new risk prediction models are generally based on the concept of events-per-variable (EPV). What this means is that to

Table 1 Considerations for developing a multivariable risk prediction model

Issue	Comment
Systematic review	Before developing a new model, carry out a systematic review to identify any existing models. Evaluate these models first to see if any are of use
Study design	Ideally use a prospective cohort study
Study population	Ensure the data used to derive the model are representative of the target population
Outcome	Logistic regression is useful for short-term outcomes (e.g., 30-day mortality). Cox regression should be used for long-term outcomes (e.g., 5-year risk)
Predictors	Only include predictors that are known or are plausibly related to the outcome that is to be predicted. Avoid screening variables on their univariate association with the outcome. Do not categorize continuous predictors. Examine nonlinearity using fractional polynomials or restricted cubic splines
Sample size	For model development, aim for a minimum of 10 events-per-variable. More events-per-variable is preferable, if possible. For model validation, aim for a minimum of 100 outcome events
Missing data	A prospective cohort design can minimize missing data. If you cannot use such a design, examine missingness and consider multiple imputation
Model presentation	Present the full model, all of the regression coefficients, and the intercept or baseline survival at one or more time points so others can use and evaluate the model and make individual predictions
Internal validation	Avoid randomly splitting data into development and validation datasets. Use bootstrapping or cross-validation to quantify and adjust for any optimism bias
External validation	Evaluate the final, published prediction model on an independent dataset
Performance	Assess both calibration and discrimination. Consider clinical utility using decision curve analysis
Comparison	Compare the prediction model with any existing models
Reporting	Follow the TRIPOD statement when writing any publications to ensure all key details of the development and validation are clearly reported

reduce the risk of overfitting, whereby the model performs optimistically well on the dataset used to develop the model, but poorly on other data, the investigator should control the ratio of number of outcome events to the number of variables examined. More appropriately, it is the number of coefficients estimated, for example, a categorical predictor with k categories, this would require k-1 regression coefficients to be estimated. Furthermore, it is the number of variables examined prior to any variable selection, including any univariate screening of individual variables, which should be avoided [46].

A minimum value of 10 EPV is widely used [47, 48] as the value to avoid overfitting in development studies, although the regression coefficients may then need shrinking. However, much larger EPV values are preferable [49, 50]. The minimum sample size recommended for validation studies is 100 outcome events. Two hundred outcome events are preferred to ensure accurate estimation of model performance [51–53].

To develop the EuroSCORE model, the authors randomly split the dataset into two cohorts, in a seemingly 90:10 ratio for the development and validation cohorts. The development dataset therefore comprised 13,302 patients, whilst the validation cohort comprised 1479 patients [23]. Neither the number of deaths nor the mortality rate was reported separately for both cohorts but only overall (698 deaths); thus, we assume there were 628 and 79 deaths, respectively (assuming a 90:10 random split). The authors should clearly describe the number of events for each separate analysis.

Missing Data

Although almost all studies are missing information in their predictors or their outcomes, missing data are often handled inadequately [54]. Missing data are often handled using a ‘complete-case’ analysis, which only includes cases with complete information on all predictors and outcomes in the analysis. However, simply excluding individuals with any missing values can lead to biased estimates and standard errors if the missing data are related to the outcome (missing at random; MAR) [55, 56]. This approach also makes the strong assumption that the reason for the missing data is not related to the outcome, which is rarely met.

Imputation approaches are more effective than complete case analysis. These approaches replace missing values from an estimate of the distribution of the observed data and assume the MAR mechanism. Single or multiple imputation can be conducted [57, 58]. Single imputation uses only one estimate (e.g., overall mean estimation) and commonly results in an underestimated standard error [59, 60]. In multiple imputation, several plausible datasets are created, and an analysis runs on each dataset. The results are combined into a single estimate with standard errors reflecting the uncertainty with the missing values. Multiple imputation leads to more correct standard errors [59, 61]. Five or 10 imputed datasets are commonly used. However, recently published rule-of-thumb recommendations suggest that the number of imputations should be larger or equal to the fraction (%) of the missing data [58]. Practical guidance for handling missing data when developing and validating risk prediction models should be followed [62–64].

In the EuroSCORE study, the handling of missing data is somewhat unclear. What the observant reader may have already noticed, is that the original EuroSCORE database comprised 19,030, yet the development ($n = 13{,}302$) and validation ($n = 1497$) sample size was substantially smaller, with and unexplained missing 4231 patients (20 % of the original dataset). With regards to completeness of individual risk factors, neither publication mentions the presence of the missing data [22, 23]. Were the 4231 patients omitted due to missing data? Is there anything special about these omitted patients? Regardless, omitting such a large proportion of the data is worrying and little is described as to why and the implications of doing so. Studies should clearly report the flow of participants, describing the missingness data at individual predictor levels as well as overall.

Modelling Continuous Predictors

Predictors are often recorded as continuous measurements, but are commonly converted into two or more categories for analysis [65]. This categorization of continuous predictors has several disadvantages. Categorisation discards information, a problem at its most severe when the predictors are dichotomized (divided into two categories). This information loss can result in a loss of statistical power and can force an incorrect relationship between the predictor and outcome. If the cut points to create the categories are not predefined, but are chosen to find the smallest P value, then the predictive performance of the model will be overly optimistic [66, 67•]. Even with a prespecified cut point, dichotomisation has been shown to be statistically inefficient [68–72]. Although using more categories reduces the information loss, this is rarely done in practice. Regardless of the number of categories, the statistical power is reduced and precision suffers in comparison with a continuous modelling approach [73]. Categorizing continuous predictors ultimately leads to poor models, as it forces an unrealistic, biologically implausible, incorrect (step) relationship onto the predictor and discards information.

The most popular approach for maintaining the continuous nature of predictors is to model a simple linear relationship between the predictor and outcome. This is often, but not always, sufficient. It may lead to a model that does not include the relevant predictor or that has an assumed relationship between the predictor and outcome that is substantially different from the "true" relationship. A better fit can be achieved using methods such as fractional polynomials (FP) or restricted cubic splines [40, 73–75]. Both of these methods allow for a nonlinear, but smooth, predictor–outcome relationship, and there is little to choose between them [67•]. Both methods can easily be implemented using standard software. FPs allow simultaneous model selection and FP specification. The results of both methods can be graphically presented, although FP results are particularly easily interpretable. It is possible to categorize predictors to implement the model if this is deemed necessary. Importantly, categorizing predictors for implementation does not require the predictors to be categorized prior to model development [76].

In the EuroSCORE study, continuous predictors were categorized using fractional polynomials [23]. It is not entirely clear what this entailed, fractional polynomials are used to describe nonlinear associations with a predictor and the outcome, and not for categorizing [73]. Nevertheless, categorizing leads to models with lower predictive accuracy [67•], and if a simple easy-to-use model is required, then more methodologically robust approaches are available [76].

Model Development

More variables are often collected than can reasonably be included in a prediction model, and therefore a smaller number of variables must be selected. Variables can be reduced before modelling by, for example, critically considering the literature, soliciting input from experts, examining correlated predictors and only including one of them, removing variables with high amounts of missing data (as these will likely be missing at the point of implementing the model in practice), and removing variables that are expensive to measure [77]. Variables are often chosen for inclusion in multivariable modelling using univariate (unadjusted) associations with the outcome. However, this common approach should be avoided as important predictors can be omitted due to confounding by other predictors [46].

Data-driven approaches such as stepwise methods (e.g., forward or backward) are common and are implemented in most statistical software. The backward selection approach is generally preferred as it considers the full model and allows the effects of all of the candidate predictors to be judged simultaneously [49]. However, these stepwise methods all have limitations in small datasets [78, 79]. When datasets are small in relation to the number of predictors examined, overfitting becomes a nonignorable concern and predictions from the model can on average be too extreme (too low or too high). Shrinkage techniques (e.g., uniform shrinkage and LASSO) can be used to reduce overfitting. Models are penalized towards simplicity by shrinking small regression coefficients towards or to zero, omitting them from the model.

The development of EuroSCORE is slightly opaque, but it seemed to include a screening of candidate risk factors based on their univariate association with the outcome (at the $P < 0.2$ level), followed by a fitting of the remaining risk factors, and their inclusion in the final model was then based on whether they improved predictive accuracy [23]. The final step was a search of first-degree interactions significant at $P < 0.05$, but it is not clear whether all possible interactions were examined or only a subset comprising those that are clinically plausible. Given the large number of centres and different countries, no attempt seems to investigate whether clustering either at a centre or in a country would have improved the model [80]. Finally, the investigators examined 97 risk factors in total; it would seem unlikely that so many risk factors could actually be plausibly related to the outcome, particularly as most risk scores only contain a handful of predictors. The consequence of having such a large number of predictors is the risk of overfitting; 97 risk factors would imply a minimum of 970 outcome events are required in the data to develop the model.

Internal Validation

Deficiencies in the statistical analysis used to develop a prediction model, such as inappropriate handling of missing data, small data, many candidate predictors, a poor choice of predictor selection strategies (including univariate screening and stepwise regression), and categorization of continuous predictors, can lead to optimistic model performance [81]. A model's performance must therefore be adequately and unbiasedly evaluated. Evaluation can be done using a so-called internal validation. Internal validation of a prediction model refers to evaluating its performance (see the "Assessing Model Performance" section) in patients from the same population that the sample originated from.

A common approach is to randomly split the dataset into two smaller datasets. The model is derived using one of these datasets (often called the training or development dataset), then its performance is evaluated using the other dataset (often called the test or validation dataset) [30]. This split-sample approach is common, but inefficient. For small to moderately sized datasets, this approach does not use all of the available data to develop the model (making overfitting more likely) and uses an inadequately small dataset for performance evaluation [81]. For large datasets, randomly splitting the data merely creates two identical datasets, which is hardly a strong test of the model.

The preferred approach for internal validation is to use bootstrapping to quantify and adjust any optimism in the predictive performance of the developed model [82]. All model development studies should include some form of internal validation, preferably using bootstrapping, particularly if no additional external validation is performed [39••].

As noted earlier, the EuroSCORE developers randomly split their data into a development cohort and a separate validation dataset, in a 90:10 split. This approach is weak and does not constitute a strong test of the model and unlikely to have the ability to quantify any overfitting.

External Validation

After a prediction model has been corrected for optimism with internal validation procedures, it is important to establish whether it is generalizable to similar but different individuals beyond the data used to derive it. This process is often referred to as external validation [30]. The more external validation studies using data from different settings, and thus different case mixes, the more generalizable the model and the more likely it will be useful in untested settings. External validation can be carried out using data from the same centres as the development data collected at a different time (temporal validation) or can be carried out using data collected from different centres (geographic validation). Model evaluation by independent investigators is a strong test of external validation. Validation is not refitting the model on new data, nor is it repeating all of the steps in the development study. Validation applies the published model (i.e., all of the regression coefficients and the intercept of baseline survival at a given time point) to new data to obtain predictions and quantify model performance (calibration and discrimination). The recommended sample size for validation studies is a minimum of 100 outcome events, preferably 200 [51–53].

Assessing Model Performance

The aim of both internal and external model validation is to quantify a model's predictive performance [17] to indicate whether it is fit for purpose and better than any existing models [24, 83•]. Discrimination and calibration are the two key characteristics of model performance that must be assessed [84••]. Discrimination is the model's ability to distinguish between individuals with and without the outcome of interest. It is commonly estimated with the *c*-index. The *c*-index is identical to the area under the receiver-operating characteristic curve for models predicting binary endpoints (e.g., logistic regression) [85]. It can be generalized for survival models accounting for censoring (e.g., Cox regression) [86].

Calibration refers to the agreement between predictions from the model and observed outcomes. That is, if the model predicts a certain risk to develop a disease, an equivalent proportion of patients with the disease should be observed in the validation sample. Calibration is preferably assessed using calibration plots showing the relationship between the observed outcomes and predicted probabilities, using a smoothed lowess line [53, 87]. Perfect calibration therefore corresponds to a slope of 1 and intercept of 0 [88]. Intercepts greater than 0 and a slope less than 1 indicate overfitting of the model [89]. The Hosmer–Lemeshow goodness-of-fit test for binary outcome models is commonly used to evaluate model calibration [16•]; however, the test has limited ability to evaluate for calibration, and is often nonsignificant (e.g., calibrated) for small sample sizes and nearly always significant for large datasets (e.g., lack of calibration). Furthermore, the tests fails to indicate magnitude of direction of any miscalibration and as such should be avoided, in preference for calibration plots [39••]. Discrimination and calibration, and other statistical methods to evaluate model performance (such as *R*-squared, Brier score) [83•] characterize the statistical properties of a prediction model, but do not capture the clinical consequences. Approaches such as decision curve analysis and relative utility should be considered to gain insight into the clinical consequences of using the model at specific probability (treatment) thresholds [90, 91].

The EuroSCORE model was evaluated by assessing discrimination and calibration. The salient point to highlight here is the use of the Hosmer–Lemeshow test [23]. Whilst the test produced P-values larger than 0.05 which the authors indicated as good calibration, this provides no meaningful indication of how well the model is calibrated. No calibration plots were presented, and as such it is unclear whether the model under- or over-predicts, or whether there are any particular subgroup of patients where the model appears less accurate.

Reporting

Numerous systematic reviews have shown that studies describing the development or validation of a risk prediction model are often poorly reported, with key details frequently omitted from published articles. Critical appraisal and synthesis are impossible when key details about the methodology used and the results are not fully reported, making it difficult for readers to judge whether a risk prediction model has any value. When a paper presents the development of a risk prediction model, it is absolutely vital that the full model, including all regression coefficients and the intercept/baseline hazard, is either presented in the paper or in an appendix, or that a link to computer code is provided, to allow other investigators to evaluate the model. Whilst this may appear obvious, many published articles on risk prediction models deliberately or unknowingly fail to report the actual model that they have developed (e.g., FRAX [92]). A nomogram (a graphical presentation of a risk prediction model [93]) is not a replacement for presenting the actual risk prediction mode [94].

In an effort to harmonize and improve the reporting of studies developing or validating risk prediction models, the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) Initiative produced the TRIPOD Statement, which is similar to the CONSORT statement for randomized clinical trials. Published simultaneously in 11 journals, the TRIPOD Statement is a checklist of 22 key items that authors should address in their articles describing the development or validation or a risk prediction model [39••, 84••].

The brief snippets we have emphasized throughout this article of specific issues in the publications describing the development of EuroSCORE highlight the problems with incomplete reporting. Only with full and transparent reporting can readers critically appraise the methodology and interpret the results.

Conclusion

Risk prediction models have great potential to aid in operative risk assessment. However, for these models to have any chance of being useful for clinical decision making, they must be developed using appropriate statistical methods and validated by others in different settings to determine their predictive accuracy. As studies developing prediction models are unfortunately rarely prospective, investigators face challenges such as how to handle missing data, what to do with continuous predictors, how to carry out an internal validation, and how to conduct a meaningful external validation study. They must also ensure complete and comprehensive reporting of every step of the study.

The surgical and anaesthesiology literature contains hundreds, if not thousands, of models developed for operative risk assessment. Only a very small minority have made any kind of impact on clinical practice. Point-and-click statistical software has arguably contributed to the plethora of methodologically weak, unusable prediction models [28]. It is therefore important to engage with a suitably experienced statistician before developing a new prediction model, to check whether a suitable model already exists, and to plan and, if possible, publish a protocol outlining the necessary steps for model development [29•].

Prediction models are usually static, reflecting the case mix in the data used to develop them. However, as mortality following surgery decreases and the case mix evolves over time, prediction models can become outdated and less accurate. This process is called calibration drift [95, 96]. Developed in 1999 using data from a 3-month period in 1995, EuroSCORE is a classic example of calibration drift [97]. The updated EuroSCORE II was therefore developed in 2012 [1], although still with some methodological concerns [98, 99]. Unless periodic updating is done, it is likely that this model will also quick become outdated.

In summary, risk prediction modelling is a growing field that is gaining huge interest in the era of personalized medicine. Although there are no shortcuts and many challenges when developing and validating accurate, useful prediction models, these challenges are surmountable, if the abundant methodological and practical guidance available is used correctly and efficiently.

Funding Jennifer De Beyer has received research funding through a Grant from Cancer Research UK.

Compliance with Ethical Guidelines

Conflict of Interest Authors declares that they have no conflict of intrerst.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as

- • Of importance
- •• Of major importance

1. Nashef SA, Roques F, Sharples LD, et al. EuroSCORE II. Eur J Cardiothorac Surg. 2012;41(4):734–44 **discussion 744–735**.
2. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375.
3. Collins GS, Jibawi A, McCulloch P. Control chart methods for monitoring surgical performance: a case study from gastro-oesophageal surgery. Eur J Surg Oncol. 2011;37:473–80.
4. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
5. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604.
6. Le Manach Y, Collins G, Rodseth R, et al. Preoperative score to predict postoperative mortality (POSPOM): derivation and validation. Anesthesiology. 2016;124(3):570–9.
7. Gurm HS, Seth M, Kooiman J, Share D. A novel tool for reliable and accurate prediction of renal complications in patients undergoing percutaneous coronary intervention. J Am Coll Cardiol. 2013;61:2242–8.
8. Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. Br J Surg. 1991;78:355–60.
9. Vuylsteke A, Pagel C, Gerrard C, et al. The papworth bleeding risk score: a stratification scheme for identifying cardiac surgery patients at risk of excessive early postoperative bleeding. Eur J Cardiothorac Surg. 2011;39:924–30.
10. Wessler BS, Lai YhL, Kramer W, et al. Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. Circ Cardiovasc Qual Outcomes. 2015;8(4):368–75.
11. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: a systematic review. BMJ. 2016;353:i2416.
12. Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol 2016;214(1):79–90 e36.
13. Shariat SF, Karakiewicz PI, Margulis V, Kattan MW. Inventory of prostate cancer predictive tools. Curr Opin Urol. 2008;18:279–96.
14. Ettema RG, Peelen LM, Schuurmans MJ, Nierich AP, Kalkman CJ, Moons KG. Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. Circulation. 2010;122(7):682–9 **687 p following p 689**.
15. Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9(5):e1001221.
16. • Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40. *Provides an overview of the conduct and reporting of external validation studies.*
17. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9:103.
18. Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney were poorly reported and often developed using inappropriate methods. J Clin Epidemiol. 2013;66:268–77.
19. Nayak S, Edwards DL, Saleh AA, Greenspan SL. Performance of risk assessment instruments for predicting osteoporotic fracture risk: a systematic review. Osteoporos Int. 2014;25(1):23–49.
20. Altman DG. Prognostic models: a methodological framework and review of models for breast cancer. Cancer Investig. 2009;27(3):235–43.
21. Collins GS, Michaëlsson K. Fracture risk assessment: state of the art, methodologically unsound, or poorly reported? Curr Osteoporos Rep. 2013;10:199–207.
22. Roques F, Nashef SAM, 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;15:816–23.
23. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg. 1999;16(1):9–13.

24. Collins GS, Moons KGM. Comparing risk prediction models. BMJ. 2012;344:e3186.
25. Masconi KL, Matsha TE, Erasmus RT, Kengne AP. Recalibration in validation studies of diabetes risk prediction models: a systematic review. Int J Stat Med Res. 2015;4:347–69.
26. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol. 2008;61(11):1085–94.
27. • Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744. *Provides a framework and gives guidance for conducting systematic reviews of prediction model studies.*
28. Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). Urology. 2010;76(6):1298–301.
29. • Peat G, Riley RD, Croft P, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med. 2014;11:e1001671. *An article stressing the importance of planning prediction model studies and if possible to register the study and publish the study protocol.*
30. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453–73.
31. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691–8.
32. Moons KG, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98:683–90.
33. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
34. Tripepi G, Heinze G, Jager KJ, Stel VS, Dekker FW, Zoccali C. Risk prediction models. Nephrol Dial Transplant. 2013;28(8):1975–80.
35. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35:1925–31.
36. Hickey GL, Blackstone EH. External model validation of binary clinical risk prediction models in cardiovascular and thoracic surgery. J Thorac Cardiovasc Surg. 2016. doi:10.1016/j.jtcvs.2016.04.023.
37. Kattan MW, Hess KR, Amin MB, et al. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. CA Cancer J Clin. 2016. doi:10.3322/caac.21339.
38. Wynants L, Collins GS, van Calster B. Key steps and common pitfalls in developing and validating risk models: a review. BJOG Int J Obst Gynaecol. 2016 (in press).
39. •• Moons KGM, 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(1):W1–W73. *Provides methodological guidance on information to report when publishing a prediction model study.*
40. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. New York: Springer; 2015.
41. • Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009. *Key text book on various aspects on prediction modelling.*
42. Groenwold RH, Moons KG, Pajouheshnia R, et al. Explicit inclusion of treatment in prognostic modelling was recommended in observational and randomised settings. J Clin Epidemiol. 2016. doi:10.1016/j.jclinepi.2016.03.017.
43. Cheong-See FM, Allotey J, Marlin N, et al. Prediction models in obstetrics: understanding the treatment paradox and potential solutions tothe threat it poses. BJOG Int J Obst Gynaecol. 2016;123:1060–4.
44. Biesheuvel CJ, Vergouwe Y, Oudega R, Hoes AW, Grobbee DE, Moons KG. Advantages of the nested case-control design in diagnostic research. BMC Med Res Methodol. 2008;8:48.
45. Sanderson J, Thompson SG, White IR, Aspelund T, Pennells L. Derivation and assessment of risk prediction models using case-cohort data. BMC Med Res Methodol. 2013;13:113.
46. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49(8):907–16.
47. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. 2. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48(12):1503–12.
48. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.
49. Ogundimu EO, Altman D, G., Collins GS. Simulation study finds adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. 2016 (in press).
50. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64:993–1000.
51. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35:214–26.
52. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58(5):475–83.
53. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76. doi:10.1016/j.jclinepi.2015.12.005.
54. Masconi KL, Matsha TE, Echouffo-Tcheugui JB, Erasmus RT, Kengne AP. Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review. EPMA J. 2015;6(1):7.
55. van der Heijden GJMG, Donders ART, Stijnen T, Moons KGM. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol. 2006;59(10):1102–9.
56. Vach W. Some issues in estimating the effect of prognostic factors from incomplete covariate data. Stat Med. 1997;16:57–72.
57. Little RJA, Rubin DB. Statistical analysis with missing data, vol. 2nd. Hoboken, NJ: Wiley; 2002.
58. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30(4):377–99.
59. Moons KG, Donders RA, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092–101.
60. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087–91.
61. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

62. Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63(2):205–14.
63. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol. 2010;10:112.
64. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;9:57.
65. Turner EL, Dobson JE, Pocock SJ. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. Epidemiol Perspect Innov. 2010;7:9.
66. Altman DG. Problems in dichotomizing continuous variables. Am J Epidemiol. 1994;139:442.
67. • Collins GS, Ogundimu EO, Cook JA, Le Manach Y, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016. *An article illustrating the loss of predictive accuracy when continuous measurements are categorised.*
68. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25:127–41.
69. van Walraven C, Hart RG. Leave 'em alone: why continuous variables should be analyzed as such. Neuroepidemiology. 2008;30:138–9.
70. Vickers AJ, Lilja H. Cutpoints in clinical chemistry: time for fundamental reassessment. Clin Chem. 2009;55:15–7.
71. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12:21.
72. Dawson NV, Weiss R. Dichotomizing continuous variables in statistical analysis: a practice to avoid. Med Decis Mak. 2012;32:225–6.
73. Royston P, Sauerbrei W. Multivariable model building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables. Chichester: Wiley; 2008.
74. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat. 1994;43(3):429–67.
75. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med. 2007;26(30): 5512–28.
76. Sullivan LM, Massaro JM, D'Agostino RB Sr. Presentation of multivariate data for clinical use: the Framingham study risk score functions. Stat Med. 2004;23:1631–60.
77. Seel RT, Steyerberg EW, Malec JF, Sherer M, Macciocchi SN. Developing and evaluating prediction models in rehabilitation populations. Arch Phys Med Rehabil. 2012;93(Suppl 2):S138–53.
78. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52(10):935–42.
79. van Buuren S, Oudshoorn CGM: Multivariate imputation by chained equations: MICE V1.0 User's Manual, vol. PG/VGZ/00.038. Leiden: TNO Preventie en Gezonheid; 2000.
80. Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Med Res Methodol. 2013;13:19.
81. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–81.
82. Harrell FE Jr, 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.
83. • Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21(1):128–138. *Paper describing many of the key performance measures to calculate when validating a prediction model.*
84. •• 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(1):55–63. *Key paper on issues to report when publishing a study developing or validating a prediction model.*
85. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008;54(1):17–23.
86. Pencina MJ, D'Agostino RB, Song L. Quantifying discrimination of Framingham risk functions with different survival C statistics. Stat Med. 2012;31:1543–53.
87. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33(3):517–35.
88. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Mak. 1993;13(1):49–58.
89. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45:562–5.
90. 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.
91. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc A. 2009;172:729–48.
92. Kanis JA, Oden A, Johnell O, et al. The use of clinical risk factors enhances the performance of BMD in the prediction of hip and osteoporotic fractures in men and women. Osteoporos Int. 2007;18(8):1033–46.
93. Kattan MW, Marasco J. What is a real nomogram? Semin Oncol. 2010;37(1):23–6.
94. Collins GS. How can I validate a nomogram? Show me the model. Ann Oncol. 2015;26:1034–5.
95. Hickey GL, Grant SW, Caiado C, et al. Dynamic prediction modeling approaches for cardiac surgery. Circ Cardiovasc Qual Outcomes. 2013;6:649–58.
96. Strobl AN, Vickers AJ, Van Calster B, et al. Improving patient prostate cancer risk assessment: moving from static, globally-applied to dynamic, practice-specific risk calculators. J Biomed Inform. 2015;56:87–93.
97. Hickey GL, Grant SW, Murphy GJ, 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.
98. Collins GS, Altman DG. Design flaws in EuroSCORE II. Eur J Cardiothorac Surg. 2012;43:871.
99. Collins GS, Altman DG. Calibration of EuroSCORE II. Eur J Cardiothorac Surg. 2013;43(3):654.