REVIEW ARTICLE

Basic principles of risk score formulation in medicine

Rafael Ronsoni[1,2]
Bruna Predabon[2]
Tiago Leiria[1]
Gustavo de Lima[1]

1. Instituto de Cardiologia do Rio Grande do Sul/Fundação Universitária de Cardiologia, Porto Alegre, RS, Brasil.
2. Universidade da Região de Joinville, Joinville, SC, Brasil.

http://dx.doi.org/10.1590/1806-9282.66.4.516

SUMMARY

Risk models play a vital role in monitoring health care performance. Despite extensive research and the widespread use of risk models in medicine, there are methodologic problems. We reviewed the methodology used for risk models in medicine. The findings suggest that many risk models are developed in an ad hoc manner. Important aspects such as the selection of risk factors, handling of missing values, and size of the data sample used for model development are not dealt with adequately. Methodologic details presented in publications are often sparse and unclear. Model development and validation processes are not always linked to the clinical aim of the model, which may affect their clinical validity. We make some suggestions in this review for improving methodology and reporting.

KEYWORDS: *Logistic models. Risk assessment. Methodology. Medicine.*

INTRODUCTION

Risk models play a vital role in the monitoring of healthcare performance and in health care policies. For a risk model to be used routinely in practice, the modeling methodology should be correct and robust. Furthermore, the proposed model must be straightforward to implement and clinically relevant[1].

Some authors provide sparse details about their methods, thus making it difficult to ascertain what was really done. Moreover, different conclusions may be reached depending on the risk model used. Therefore, it is important to ensure that risk modeling is carried out in a correct and systematic manner and that robust and accurate models are developed. Since some deficiencies may exist in the earlier processes, thus making the clinical application questionable, the objective of the risk model should be clear to prevent it from being used in the same way as other clinical situations previously studied. The performance of a model should be evaluated in light of the specified goals[2,3].

The objective of this study is to review the methodology used for risk modeling in medicine, suggesting a correct methodological sequence for avoiding common errors in the development of this type of research.

STEPS

Objectives of the risk score

The first step is the formulation of clear objectives, as these will have effects on the choice of variables

DATE OF SUBMISSION: 28-Oct-2019
DATE OF ACCEPTANCE: 12-Nov-2019
CORRESPONDING AUTHOR: Rafael Ronsoni
Rua Duvoisin, 3, Joinville, SC, Brasil - 89204-358
Tel: +55 47 99163-4050
E-mail: rafaelronsoni@gmail.com

to be studied and be directly involved in the clinical application of the model[4]. (figure 1 and 2)

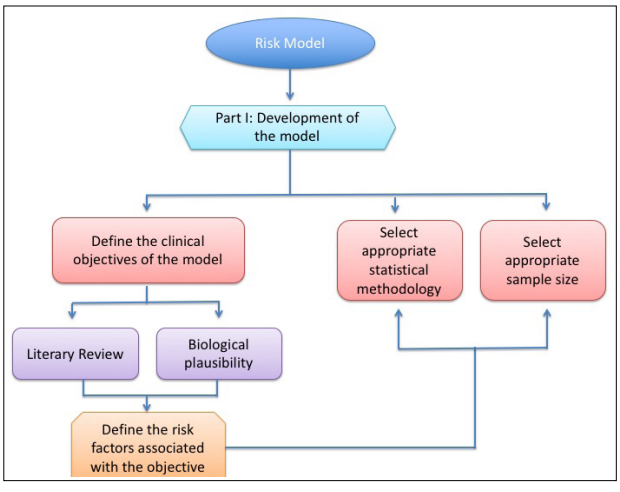
Choice of variables to be studied

The choice of variables usually follows a hierarchical model based on biological plausibility and external information (ie., literature) regarding the strength of the associations (along with the occurrences) related to the study outcome; associations are of fundamental importance[3,5]. The choices of the variables are in accordance with the clinical objectives specified for the model in question. Likewise, they demonstrate a balance between the complexity required and what can be collected in clinical practice. Highly complex risk models can satisfy the most diverse clinical objectives, but may be impractical and even despised in clinical practice[4]. Parsonnet et al.[6] suggests testing the evaluated variables that have a prevalence greater than 2% in the sample to avoid possible bias.

In choosing the variables, we tried to minimize bias in this detailed situation. The value of the regression coefficient was calculated (see below) to be as accurate as the average effect of X, but the result would be misleading if X has different effects in different zones. The implication may be particularly misleading if the average value of X does not occur in any of the zones. For example, the impact of left ventricular ejection fraction on mortality is not linear: A decrease of 10%, from 30% to 20%, carries greater risk than a decrease from 50% to 40%[7].

An important detail at the end of the score is that all the risk factors surveyed were generated in the final model are presented[4].

FIGURE 1



Definition of derivation cohort or development group and their size

The first analysis usually occurs with a specific sample of patients. This is called a derivation cohort or developmental group, which is basically the primary objective in the development of the prediction score[3].

The number of events per variable analyzed by logistic regression should be greater than or equal to 10, minimizing possible statistical errors[8]. In general, the results of models with less than 10 outcome events per independent variable are thought to have questionable accuracy, and the usual tests of statistical significance may be invalid. Large confidence intervals associated with individual risk estimates may indicate an over-fitted model under these circumstances[7].

Application of logistic regression analysis and preliminary score

After the variables in the sample were studied, multiple logistic regression was applied. The predictor score was then calculated. This was derived by utilizing the variables that are true and independent factors, generally in keeping with the score of all the variables with a level of significance of $p < 0.05$[3-5]. Generally, the variable selection strategy employed by the score developers to produce their final score used variable selection methods such as backward elimination, forward selection, and stepwise approaches[9].

Typically, a risk score produces coefficients for each risk factor in the final score representing their weights in predicting studied outcome. If not chosen this way, there are methods available to translate coefficients into integer scores with minimal loss of precision[10]. One of the more commonly used forms is a risk-weighted score based on the magnitude of the coefficients b of the logistic equation. When they were transformed (exp [b]) into odds ratios (odds ratios), the values were rounded to make up the initial predictor score[11].

Score Predictor calibration test

Ideally, the prediction score should be subjected to calibration and discrimination tests. Calibration evaluates the accuracy to predict risk in a group of patients. More succinctly, if the score proposes that the clinical event in 1,000 patients would be 5% and the observed clinical event is 5% or close to that value, it would be prudent to conclude that the model is well-calibrated. The strength of the calibration can be assessed by testing the quality of the fit using the Hosmer-Lemeshow

test[12,13]. A p value> 0.05 indicates that the score fits the data and predicts the outcome adequately.

Score Predictor discrimination test

Discrimination measures the ability of the score to distinguish between low-risk and high-risk patients. In other words, if most clinical events occur in patients identified as high risk, we will say that the model has good discrimination. On the contrary, if most clinical events occur in patients identified as low risk by the model, we will say that the model has poor discrimination[14].

Discrimination is measured using the statistical technique called area below the ROC curve. It is typically used for the evaluation of prognostic models in cardiology and represents the likelihood of a predictive model. In the case of a risk score, it is used to assign a higher probability of an event occurring in those who will actually present the event. The area under the ROC curve is a summary of the accuracy of the score and is represented by the C (concordance) statistic for binary outcomes. C is equal to 0.5 when the ROC curve corresponds to the probability, represented by the diagonal line in the curve, and results in 1.0 when the accuracy is maximum in discriminating between those with and without the outcome under study. For example, a ROC area of 0.75 means the model correctly ranks 75% of the patient pairs according to their predicted probability. Risk score with statistic C classified as excellent discrimination refers to values above 0.97; very good discrimination is in the range of 0.93 to 0.96; good discrimination between 0.75 and 0.92; below 0.75 corresponds to models deficient in the ability to discriminate. In practice, models rarely exceed 0.85[5,13,15,16].

Internal validation of the Score Predictor in a validation cohort

Like all prediction scores, the initial score needs to be validated (figure 3). The evaluation of the performance of the prediction model in data not belonging to the derivation cohort is the most important. This can be achieved using internal validation, which is the submission of the model to a new population of the same center and evaluating its predictive performance in the second group, the so-called validation cohort[3].

Depending on the aims, the validation process should consider the total picture; the ability of a model to predict the outcome of the risk score accurately; the range of predictions (whether these are clinically useful or not); and the ability to discriminate between high, intermediate, and low-risk patients[4].

The validation dataset should be large enough to enable precise comparison between the outcomes observed and predicted and to enhance statistical methods such as the H-L test with sufficient power. For the H-L test to be valid, the predicted number of events in each risk group used in the test should always be greater than 1, and for most risk groups it should be at least 5[12]. It has been suggested as a general principle that adequate model evaluation requires at least 100 outcomes in the validation sample[9].

Logistical Model

In addition to the final score, the resulting logistic model can be presented (see formula below), in which it is possible to obtain direct estimates of the probability of occurrence of an outcome. This process, using the mathematical model directly, is understood and regarded by some authors as the most appropriate in obtaining event estimates, although it presents a

FIGURE 2

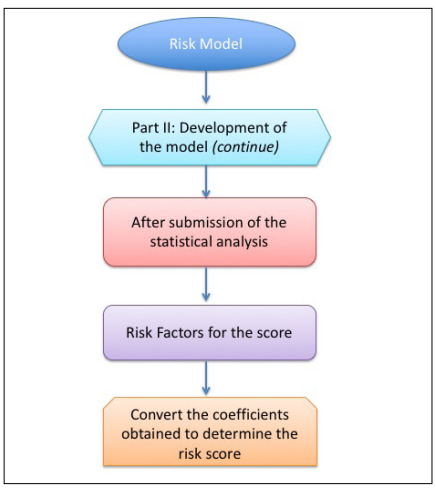
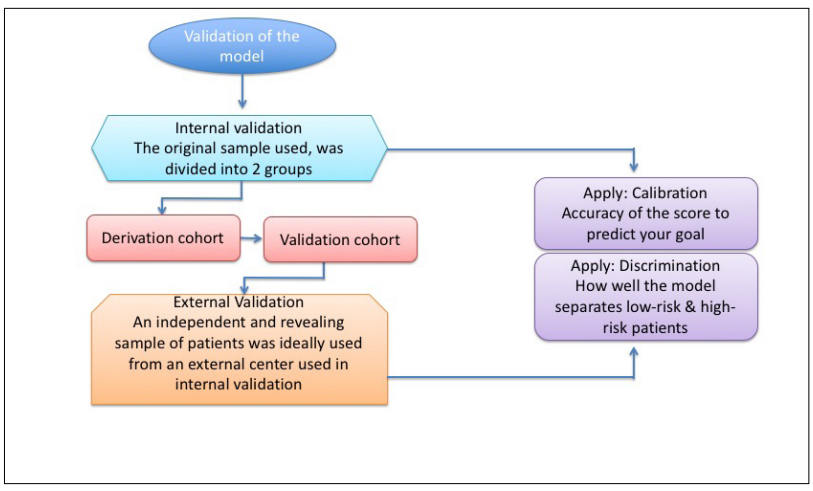


FIGURE 3



certain degree of mathematical complexity for its use in daily medical practice. The application of the logistic model is more adequate for the prognosis of individual risk, especially in patients with very high risk in the additive model[17].

$$P(event) = 1 / 1 + \exp(- (\beta 0 + \beta 1x1 + \ldots + \beta kxk))$$

LIMITATIONS OF ANALYSIS OF RISK SCORES

The analysis of the performance of a risk model based only on its discriminatory capacity (C-statistics) and calibration has limitations. One of the main limitations to be highlighted is the observation that once the area under the ROC curve reaches a certain level, large sizes of new variable effects are required to achieve small increases in the area under the ROC curve. Due to these limitations, new methods of quantifying performance improvement have been developed, such as risk reclassification, Net Reclassification Improvement (NRI), and Integrated Discrimination Improvement (IDI)[18].

Several studies suggest that the scores are less effective when applied to patients outside the scope of the target group intended for the study. Therefore, external validation is fundamental to increase its clinical acceptance, especially for centers outside the site of the creation of the score[3,4]. As with any risk stratification score, it should always be evaluated and re-evaluated in the long term, considering existing variables. Risk score methodologies should also be designed to accommodate and incorporate the presentation of new variables.

CONCLUSION

Risk models play an important role in health care policy. For a risk model to be used routinely in practice, the modeling methodology should be correct and robust. The proposed model must be straightforward to implement and clinically relevant. It is important that researchers implement a structured and transparent model-building process linked to the stated clinical objective. Furthermore, it is imperative they evaluate the model's performance in the context for which it had been developed. Researchers should also develop guidelines before starting a modeling process, describing how each step will be handled, and strictly adhere to these protocols. Clearer descriptions of each step of the process are required in published papers and reports.

Authors contribuition

Rafael Ronsoni: Conceptualization (Equal), Data curation (Equal), Formal analysis (Equal), Funding acquisition (Equal), Investigation (Equal), Methodology (Equal), Project administration (Equal), Resources (Equal), Software (Equal), Supervision (Equal), Validation (Equal), Visualization (Equal), Writing-original draft (Equal), Writing-review & editing (Equal)

Bruna Predabon: Conceptualization (Equal), Data curation (Equal), Formal analysis (Equal), Funding acquisition (Equal), Investigation (Equal), Methodology (Equal), Project administration (Equal), Resources (Equal), Software (Equal), Supervision (Equal), Validation (Equal), Visualization (Equal), Writing-original draft (Equal), Writing-review & editing (Equal)

Tiago Leiria: Conceptualization (Equal), Data curation (Equal), Formal analysis (Equal), Funding acquisition (Equal), Investigation (Equal), Methodology (Equal), Project administration (Equal), Resources (Equal), Software (Equal), Supervision (Equal), Validation (Equal), Visualization (Equal), Writing-original draft (Equal), Writing-review & editing (Equal)

Gustavo de Lima: Conceptualization (Equal), Data curation (Equal), Formal analysis (Equal), Funding acquisition (Equal), Investigation (Equal), Methodology (Equal), Project administration (Equal), Resources (Equal), Software (Equal), Supervision (Equal), Validation (Equal), Visualization (Equal), Writing-original draft (Equal), Writing-review & editing (Equal)

RESUMO

Os modelos de risco desempenham um papel fundamental no monitoramento dos desempenhos dos serviços de saúde. Apesar de extensa pesquisa e do amplo uso dos modelos de risco na Medicina, existem problemas metodológicos. Revisamos a metodologia utilizada nestes modelos na Medicina. Os achados sugerem que muitos modelos de risco são desenvolvidos de maneira ad-hoc. Aspectos importantes, como a seleção de fatores de risco, a forma utilizada de dados perdidos e o tamanho da amostra empregada não são detalhados adequadamente. Detalhes metodológicos presentes em publicações são frequentemente esparsos e incertos. Os modelos de desenvolvimento e de validação nem sempre estão associados com o objetivo clínico do modelo, o que pode afetar sua validade clínica. Nós produzimos algumas sugestões nesta revisão para otimizar a metodologia e as publicações.

PALAVRAS-CHAVE: *Modelos logísticos. Medição de risco. Metodologia. Medicina.*

REFERENCES

1. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med. 1993;118(3):201-10.
2. Omar RZ, Ambler G, Taylor KM. Assessment of quality of care in cardiac surgery: an overview of risk models. In: Mohan R, eds. Recent advances in cardiology. Global Research Network. India [Internet]. 2003 p.13-20. [cited 2019 Jun 21]. Available from: http://discovery.ucl.ac.uk/77344/
3. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453-73.
4. Omar RZ, Ambler G, Royston P, Eliahoo J, Taylor KM. Cardiac surgery risk modeling for mortality: a review of current practice and suggestions for improvement. Ann Thorac Surg. 2004;77(6):2232-7.
5. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
6. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation. 1989;79(6 Pt 2):I3-12.
7. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Intern Med. 1993;118(3):201-10.
8. 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.
9. Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001. 571p.
10. Cole TJ. Algorithm AS 281: scaling and rounding regression coefficients to integers. Appl Stat [Internet]. 1993;42(1):261-8. [cited 2019 Jun 18]. Available from: https://www.jstor.org/stable/2347432?origin=crossref
11. Guaragna JCVC, Bodanese LC, Bueno FL, Goldani MA. Proposta de escore de risco pré-operatório para pacientes candidatos à cirurgia cardíaca valvar. Arq Bras Cardiol. 2010;94(4):541-8.
12. Hosmer DW, Lemeshow S. Applied logistic regression. Hoboken: John Wiley & Sons Inc.; 2000 [cited 2017 Jan 12]. Available from: http://doi.wiley.com/10.1002/0471722146
13. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol. 1982;115(1):92-106.
14. Jones CM, Athanasiou T. Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. Ann Thorac Surg. 2005;79(1):16-20.
15. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007;115(5):654-7.
16. LaValley MP. Logistic regression. Circulation. 2008;117(18):2395-9.
17. Zingone B, Pappalardo A, Dreas L. Logistic versus additive EuroSCORE. A comparative assessment of the two models in an independent population sample. Eur J Cardiothorac Surg. 2004;26(6):1134-40.
18. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157-72.