



# How to use statistical models and methods for clinical prediction

Giuliana Cortese

Department of Statistical Sciences, University of Padova, Padova, Italy

Correspondence to: Giuliana Cortese, Department of Statistical Sciences, University of Padova, Via C. Battisti 241, 35121 Padua, Italy.

Email: gcortese@stat.unipd.it.

Provenance: This is an invited article commissioned by the Editorial Office of *Annals of Translational Medicine*.

Comment on: Zhou ZR, Wang WW, Li Y, *et al.* In-depth mining of clinical data: the construction of clinical prediction model with R. *Ann Transl Med* 2019;7:796.

Submitted Dec 21, 2019. Accepted for publication Jan 03, 2020.

doi: 10.21037/atm.2020.01.22

View this article at: <http://dx.doi.org/10.21037/atm.2020.01.22>

One of the main aims of statistics is to control and model variability in observed phenomena. A second important aim is to translate the results of such modelling into clinical decision-making, e.g., by constructing appropriate prediction models. Currently, model-based individualized predictions play an important role in the era of personalized medicine, where diagnosis and prognosis of a clinical outcome are based on a large number of observed clinical, individual and genetic characteristics (1). The paper by Zhou *et al.* (2) describes an interesting summary of clinical prediction models that range from the establishment of a clinical problem, study design and data collection to the identification, construction, validation and assessment of the effectiveness of a prediction model. Moreover, it presents a brief discussion about the necessity to update a clinical prediction model over time and current practical issues. Finally, most of the paper is dedicated to the implementation in R of the different steps of construction, validation and effectiveness of two key examples of prediction models, the logistic regression model for categorical data and the Cox proportional hazards model for survival data (time-to-event data). The overview of how to apply the different R packages is highly useful and promotes the translation of statistical theory to its practical use.

Zhou *et al.* (2) mention that machine learning is used to build models, mostly nonparametric models. In reality, machine learning can be both parametric and nonparametric, just as there exist parametric, semi-parametric and nonparametric statistical models. Machine learning methods are created for the purpose of providing the most accurate predictions possible, and therefore

parameter interpretability is traditionally not the focus of this approach. Machine learning uses statistical models to learn from the data with the aim to make repeated predictions using “new” external data (called testing set). That is the reason why it works well and is suitable for a large amount of data.

On the other side, statistical regression models are typically constructed to make inference on the effects of clinical predictors, i.e., to infer the relationships between variables and discover insights in the underlying target population. However, statistical models can also be used to predict future outcome values and they can be compared in terms of prediction accuracy. Moreover, statistical models require an inferential and probabilistic theory with well-defined properties and can also fit a relatively small sample of data. Therefore, these two different approaches integrate each other and should not be considered as contrasting approaches when the purpose is prediction (3). Indeed, it is not necessarily true that machine learning provides an improvement in predictive performance over classical statistical methods, as shown in some practical examples (4,5). However, other clinical studies showed that machine learning is superior (6).

The authors write that in algorithms of machine learning, the clinical interpretation of nonparametric models is difficult. This is in general true whatever the statistical approach is, when there are only “functions”, rather than single-value regression coefficients, to quantify the effects of predictors on the clinical outcome. Nonparametric or semi-parametric techniques are nowadays very useful to find a valid statistical model that is free from restrictive

assumptions, which are often not fulfilled when fitting complex clinical data. In the literature there exist various semi-parametric and nonparametric extensions of survival Cox models (7).

Direct and simpler clinical interpretation of the resulting estimates is a general issue that not only concerns regression coefficients, but also involves regression model specification. A recent example of that is the open discussion in the medical scientific community about hazard ratios resulting from a fitted survival Cox model. Often the primary clinical interest is on survival probabilities or alternative summary measure over the entire follow-up. Hazard ratios are difficult to interpret, and it is hard to translate their effects into clinical benefits due to prolonged survival time (8). For these reasons, there is an increasing interest towards alternative regression models where regression coefficients quantify direct effects of predictors on survival probability, survival mean life or residual mean life (9).

The methods described in Zhou *et al.* (2) are only a few basic examples of a very extensive literature about statistical models for clinical data and related techniques for prediction performance. In regard of this, as also stressed in the paper by the authors, it is very crucial to strengthen the collaboration between clinical scientists and biostatisticians in order to choose the most appropriate statistical tools for the specific clinical problem at hand.

Concerning variable selection in regression analysis, existing statistical methods such as likelihood ratio test, Wald chi-square test, Akaike Information Criteria, etc. should not be considered as contrasting options, since they answer to different questions. These methods support each other in the complex process of variable selection, which also requires a strict collaboration between clinical researchers and statisticians, and thus we should not expect to have a “best” method for variable selection. For example, likelihood ratio test is used to compare models in terms of increased explained variability, while Wald chi-square test refers to statistical significance of a single covariate. The reason why some studies reports contrasting results about significance of covariates might be due to many reasons, e.g., background population from which the data are sampled might be different, collinearity between variables is present, statistical interactions between variables have been ignored, etc.

To study the prediction accuracy of prediction models, we should assess discrimination, calibration and be able to compare performance of different prediction models. Many indexes are nowadays available. Some of them are designed to measure the ability of the model to discriminate, e.g.,

the c-index (10) and the area under the ROC curve (11). However, these indexes do not account for calibration of the model, which is also an important aspect to ensure accurate predictions. Calibration has the purpose to measure the agreement between observed outcomes and predicted probabilities. The paper by Alba *et al.* (12) provides a very good and complete user’s guide to obtain a clinical prediction model, illustrating the application of available indexes. Moreover, the literature offers novel measures of prediction accuracy based on the proportion of explained variance to quantify the potential predictive power, and on the proportion of explained prediction error (13).

Zhou *et al.* (2) underlines the importance of external data, as well as internal data, for model validation in assessing prediction accuracy. This fact is essential to be able to translate the model predictions into the process of clinical decision making. In addition, when multiple validations are performed, a meta-analysis may help to summarize overall performance across multiple settings and sub-populations (14).

Nomograms are important tools to account for a large number of covariates in clinical predictions. An interesting and comprehensive application that shows how to derive and validate a nomogram is presented by Liu *et al.* (15). The authors discuss whether is more convenient to prefer accuracy to practicality when selecting the number of variables used to build a nomogram. This is also an important practical aspect not to be ignored and suggests again that the trade-off between maximal prediction accuracy and implementation in clinical practice should be well addressed by a collaborative working team of biostatisticians and medical doctors. On the other hand, the new era of big data encourages considering more and more available information (and variables) in future medical practice.

## Acknowledgments

*Funding:* G Cortese was supported by the National Project PRIN 2017 (Prot. 20178S4EK9), Italian Ministry of Education, University and Research.

## Footnote

*Conflicts of Interest:* The author has no conflicts of interest to declare.

*Ethical Statement:* The author is accountable for all

aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## References

1. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing, 2009.
2. Zhou ZR, Wang WW, Li Y, et al. In-depth mining of clinical data: the construction of clinical prediction model with R. *Ann Transl Med* 2019;7:796.
3. Austin PC, Tu JV, Ho JE, et al. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *J Clin Epidemiol* 2013;66:398-407.
4. Tollenaar N, van der Heijden PGM. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *J R Statist Soc A* 2013;176:565-84.
5. Christodoulou E, Ma J, Collins GS et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12-22.
6. Matsuo K, Purushotham S, Jiang B et al. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am J Obstet Gynecol* 2019;220:381.e1-14.
7. Cortese G, Scheike TH, Martinussen T. Flexible survival regression modelling. *Stat Methods Med Res* 2010;19:5-28.
8. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 2014;32:2380-5.
9. Cortese G, Holmboe SA, Scheike TH. Regression models for the restricted residual mean life for right-censored and left-truncated data. *Stat Med* 2017;36:1803-22.
10. 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.
11. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005;61:92-105.
12. Alba A, Agoritsas T, Walsh M, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 2017;318:1377-84.
13. Li G, Wang X. Prediction Accuracy Measures for a Nonlinear Model and for Right-Censored Time-to-Event Data. *J Am Stat Assoc* 2019;114:1815-25.
14. Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* 2019;28:2768-86.
15. Liu H, Zheng S, Li X, et al. Derivation and validation of a nomogram to predict in-hospital complications in children with tetralogy of fallot repaired at an older age. *JAHA* 2019;8:e013388.

**Cite this article as:** Cortese G. How to use statistical models and methods for clinical prediction. *Ann Transl Med* 2020;8(4):76. doi: 10.21037/atm.2020.01.22