



Models and prediction, how and what?

Yujing Xie¹, Zhangsheng Yu^{1,2}

¹School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200030, China; ²Department of Bioinformatics and Biostatistics, SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai 200030, China

Correspondence to: Zhangsheng Yu. Department of Bioinformatics and Biostatistics, SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, 800 Dongchuan RD. Minhang District, Shanghai 200030, China. Email: yuzhangsheng@sjtu.edu.cn.

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 08, 2019. Accepted for publication Dec 13, 2019.

doi: 10.21037/atm.2019.12.133

View this article at: <http://dx.doi.org/10.21037/atm.2019.12.133>

Prediction and estimation are the two primary goals in statistical modeling. Prediction models are essential for medical researches, especially with the rapid accumulation of medical data. Zhou *et al.* recently made a comprehensive summary of the prediction model and its evaluation methods (1). The work provides in depth tools in prediction models for medical researchers and bridges the statistical method researchers and medical applications. The application driven introduction of prediction model provides prediction model analysis workflow without involving detailed formula for the medical investigators. The included examples with R code and interpretation are illuminating for statistical practice.

The article covers the concepts of clinical prediction models, variable screening, Nomogram plotting, C-index, net reclassification index (NRI) and integrated discrimination index (IDI) calculations, outliers and missing value processing etc. in 16 sections. The rich content is worthy of recognition. Due to the wide coverage, it would be beneficial to add a directory or index section to make it easier for readers to locate the specific content.

For the evaluation metrics of the prediction model, the detailed summary of the authors provides a technical basis for the subsequent research work. For example, based on the case study, the article gives an implementation of R code for nomogram in logistic regression and Cox regression models. The concept of Nomogram was proposed by Ocagne (2) and was first used in engineering. It can display complex engineering mechanics and other calculation formulas graphically and intuitively. The first medical nomogram

was from the work of Lawrence Henderson, who used a nomogram to correlate many different aspects of blood physiology. Since then, Nomograms continue to be widely used in medical field, even in recent medical research. In the study of the preoperative prediction of lymph node metastasis in bladder cancer, Wu *et al.* (3) established a logistic regression model using radiological features extracted from the CT images of each patient's arterial phase. And the used radiomics nomogram showed good calibration and discrimination in the validation set. In addition, there are some research works based on Nomogram. Mohamed *et al.* (4) compared the predictive characteristics of the practice of plotting transcutaneous bilirubin (TcB) values on the total serum bilirubin (TSB) hour-specific risk nomogram versus on transcutaneous nomogram. They found that plotting TcB on specific transcutaneous nomograms resulted in better predictive characteristics and was associated with lower false negative results and higher positive predictive value. Due to the popularity of Nomogram in the medical field, the authors' summary of its implementation and interpretation is necessary.

In addition to summarizing the classical predictive metrics (e.g., C-index), the article also discusses new methods proposed in recent years for prediction model comparisons, such as decision curve analysis (DCA). In practice, for the selection of two prediction models, we usually think that the model with a large AUC value on the validation set is better. Vickers and Elkin (5) pointed out that AUC cannot tell us which prediction model is preferable because metrics that concern accuracy do not

incorporate information on consequences. They proposed a novel indicator for evaluating the prediction model, namely the DCA. In recent years, based on DCA, many researchers have conducted related extension studies. For example, Tsalatsanis *et al.* (6) considered that rational decision-making should reflect both formal principles of rationality and intuition about good decisions and used cognitive emotion of regret to reformulate DCA. Therefore, the authors' work about the detailed guidance on the implementation of DCA in R will help more medical researchers to use newer research results to improve the predictive model criteria, so that cutting-edge theory can be better used in clinical practice.

On the other hand, there are statistical issues worth noting. The authors mentioned that a normality test of the independent variables is required before performing logistic regression or Cox model. We know that the distribution of the covariates of the statistical model is assumed to be unknown, so this test is not actually unnecessary and may increase the difficulty of interpretation. The main model evaluation methods of the article are based on the classical logistic and Cox models, which are based on a linear assumption of the covariate effects. However, in practices, linear assumptions are often violated. Although the linear model has the advantage of interpretation, it is necessary to consider the non-parametric model when the linear condition is violated. For the non-parametric model, the R package *mgcv* based on spline estimation provide tools for fitting prediction model with satisfactory prediction accuracy. In addition, some machine learning methods such as neural network or XGBoost can also be used for nonlinear model or multi-category outcome prediction, which can be easily implemented in R software by *neuralnet* package or *xgboost* package. Because some of the methods for evaluation models' prediction ability are not dependent on the linear model assumption, such as C-index and nomogram etc., the content of the article can be further extended to more general prediction models such as non-parametric models, with some changes in the code implement.

In summary, we congratulate the authors for great work

on prediction models building guidance. Some extension in using more flexible covariate function forms and modelling tools in machine learning will be great additions.

Acknowledgments

Funding: The research was supported in part by National Natural Science Foundation of China (11671256 Yu).

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

Ethical Statement: The authors are 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. 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.
2. Ocagne P. *Traité de nomographie*. Mon Hefte Math 1900;11:A9-10.
3. Wu S, Zheng J, Li Y, et al. A Radiomics Nomogram for the Preoperative Prediction of Lymph Node Metastasis in Bladder Cancer. *Clin Cancer Res* 2017;23:6904-11.
4. Mohamed I, Blanchard AC, Delvin E, et al. Plotting transcutaneous bilirubin measurements on specific transcutaneous nomogram results in better prediction of significant hyperbilirubinemia in healthy term and near-term newborns: a pilot study. *Neonatology* 2014;105:306-11.
5. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565-74.
6. Tsalatsanis A, Hozo I, Vickers A, Djulbegovic B. A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. *BMC Med Inform Decis Mak* 2010;10:51.

Cite this article as: Xie Y, Yu Z. Models and prediction, how and what? *Ann Transl Med* 2020;8(4):75. doi: 10.21037/atm.2019.12.133