Editorial

The power of clinical data empowered by clinical prediction model: an R tutorial

Lutao Dai, Dan Yang, Haipeng Shen

Faculty of Business and Economics, The University of Hong Kong, Hong Kong, China

Correspondence to: Haipeng Shen. Faculty of Business and Economics, The University of Hong Kong, Pok Fu Lam, Hong Kong, China.
Email: haipeng@hku.hk.

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 31, 2019. Accepted for publication Jan 16, 2020.
doi: 10.21037/atm.2020.01.114
View this article at: http://dx.doi.org/10.21037/atm.2020.01.114

Clinical practice is going through disruptive transformation from physician-centered decision-making to data-driven, low-cost, and machine-generated judgement. This transformation is powered by the striking advancement of computation, memory and storage, and the unprecedented abundant resources of clinical data. Clinicians resort to not only their own instincts but also valuable information from clinical prediction models, which are constructed meticulously and sometimes perform on par with human physicians. However, building reliable clinical models that have practical values requires many assumptions and a series of steps, such as construction, evaluation, and validation. Zhou *et al.* (1) provides a timely and comprehensive users' guide for such a purpose, demonstrating the associated steps in R step-by-step.

Broadly speaking, clinical prediction models refer to those statistical models applied in clinical settings. They try to identify patients' current diseases or derive the likelihood of some future clinical events, such as death, disability, or other complications based on known personal features and historical clinical events. They provide an objective, data-driven perspective on patients' conditions, doctors' decision quality, and hospitals' operational efficiency, which helps clinicians, patients, and health management departments to improve their decision-making process. Current well-established and widely-applied prediction models include the TNM staging system (2), Framingham (3) and QRISK (4). Clinical prediction models have the potential to deliver more accurate and timely predictions, which undoubtedly have significant social values.

Zhou *et al.* (1) provides clear guidance for a complete cycle of clinical prediction model development, including selection of different model types (which in turn depends on different types of research design and different data types), data preprocessing, variable selection tools, various model evaluation methods, and model validation techniques.

Model identification

Zhou *et al.* (1) first summarized three types of clinical issues such as diagnostic models, prognostic models, and disease occurrence models. They further connected these models to cross-sectional studies and cohort studies. Depending on the nature of the outcome and problem of interest, they mainly discussed three types of statistical models: multivariate regression for quantitative response, logistic regression for binary response, and survival analysis with both Cox model and competitive risk model.

Preprocessing

Once a specific model is chosen, before parameter estimation, data preprocessing is necessary. Zhou *et al.* (1) discussed in detail on missing data, outliers, categorical variable, and variable screening. We want to add that, since the seminal work by Fan *et al.* (5), many more advanced

variable screening methods have been studied and are scalable to handle high-dimensional data and more effective than traditional methods (6-9).

Model evaluation

Once data are cleaned and the model is estimated, it is essential to evaluate its performance. For each of the possible models on classification and survival analysis, Zhou *et al.* (1) discussed at length about three indices (C-Index, Net Reclassification Index, and Concordance index of discrimination) and two visualization tools (Nomogram drawing and Decision Curve Analysis). They compared the strengths and weaknesses of these tools and demonstrated their usage in R.

Model validation

Model validation refers to evaluating both internal and external validity. The fundamental reason for validation is to deal with overfitting. The good performance on training data does not guarantee good model generalizability. Models trained on a training set has the tendency to capture the noise pattern, especially when the model is fairly flexible. Internal validity is concerned with reproducibility of the model on the same data source and is commonly assessed by cross validation or bootstrap. External validation validates the model on a dataset from different sources and should be emphasized when making general conclusions.

Statistical learning has been widely and successfully used in many fields and it is beneficial for clinical practitioners to rely on Zhou *et al.* (1) for a comprehensive and coherent introduction to the complete picture of the application of statistical learning in the field of clinical studies while focusing on the uniqueness and key differences from other fields. Machine learning is regarded as "the fundamental technology to process data to exceed the capacity of the human brain to comprehend", yet until now its application to health care is surprisingly sparse (10). Zhou *et al.* (1), along with many others (10-12), attempts to motivate more research in the intersection of machine learning and health care by informing the researchers of the general principles and considerations of developing such models. Unlike the other review papers, Zhou *et al.* (1) provides detailed R implementations to substantiate its methodological review. By following the examples, the readers will have a more concrete idea of how to achieve the goals at different stages of model development.

Beam *et al.* (12) proposes the idea of *machine learning spectrum*, which is a continuum between fully human-guided algorithms and fully machine-guided data analysis. The major models discussed by Zhou *et al.* (1), such as logistic regression and Cox regression, should be placed somewhere closer to the lower end, because they rely heavily on human experts' inputs and prior assumptions. On the contrary, the other review papers (10-12) also include a decent amount of discussion on deep learning models, a type of more flexible models capable of deriving hierarchical features by themselves, and therefore place the methodology on the higher end of the spectrum. This is understandable, since Zhou *et al.* (1) centers its discussion around R implementation, while R is not one of the popular languages for constructing deep learning architectures.

The healthcare research literature appreciates the interpretability of simpler models (13). Nevertheless, over the past few years, the field has witnessed many successful applications of deep learning models, some even reporting expert-level performance (14-16). Besides encouraging results, deep learning is a more scalable method, which is especially advantageous in the era when medical data are abundant and data sources are more diversified, such as those from electronic healthcare records (EHR), genetic testing, mobile devices, and even wearable sensors (13). It is infeasible to perform variable selection manually and difficult to determine the appropriate size of variables to be included. Additionally, traditional machine learning methods, including the ones from Zhou *et al.* (1), are not proficient at processing non-tabulated data, such as images. As a result, people generally believe that deep learning shows great promise to revolutionize the health care research. Readers are encouraged to learn more about the deep learning methodology in healthcare by referring to (10-12).

Regardless of the position in the spectrum, robust and valid results are always the key. Zhou *et al.* (1) has extensive discussion on performance metrics and how to evaluate the validity of the model. However, one important consideration left out is ethics, which may be seemingly irrelevant to the clinical prediction model development but actually a necessary aspect that should be addressed throughout the development process, especially in medical settings (17). For instance, prior to building the model,

researchers should evaluate and be alert to any potential biases inherent in the data, such as racial and geographical biases, so that their models will be less likely to reach erroneous conclusions targeting at a specific group of people. Consequentially, they can seek to mitigate the bias by collecting more diversified data or be critical about the application scenarios of the model to be developed. In the stage of assessing the model behavior, it is important to evaluate whether the model picks up human biases in decision making and produces discriminatory outcomes. Lastly, when implementing the model, patients' medical records will be inevitably exposed, even though unlikely to the public. This challenges the traditional understanding of respecting patients' privacy by withholding their records. The field will eventually have to redefine the doctors' duty of confidentiality to patients.

As mentioned in Zhou *et al.* (1), "the statistical nature of regression analysis is to find the 'quantitative causality'." To be simple, regression analysis is a quantitative characterization of how much X affects Y. It should be noted that regression models are in general only tools to identify correlation, not causation. It is well understood that causality is the ultimate goal in most clinical studies. But the tools introduced in Zhou *et al.* (1), and more so those more advanced deep learning tools introduced in the other review papers, cannot achieve such a goal without seeking help from the literature on causal inference for observational studies, unless rigorous randomized experiments could be conducted without non-adherence so that confounders would not bias the results.

Acknowledgments

We sincerely thank the Editor for inviting us to comment on the Zhou *et al.*

Funding: This work is supported by Ministry of Science and Technology Major Project of China 2017YFC1310903, University of Hong Kong (HKU) Stanley Ho Alumni Challenge Fund, HKU University Research Committee Seed Funding Award 104004215, HKU BRC Fund, and US NSF BIGDATA Grant 1741390.

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. Goldstraw P, Chansky K, Crowley J, et al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. J Thorac Oncol 2016;11:39-51.
3. Wilson PW, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837-47.
4. Hippisley-Cox J, Coupland C, Vinogradova Y, et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ 2007;335:136.
5. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Series B Stat Methodol 2008;70:849-911.
6. Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Am Stat Assoc 2012;107:1129-39.
7. Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Stat 2010;38:3567-604.
8. Wang H. Forward regression for ultra-high dimensional variable screening. J Am Stat Assoc 2009;104:1512-24.
9. Hao N, Zhang HH. Interaction screening for ultrahigh-dimensional data. J Am Stat Assoc 2014;109:1285-301.
10. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347-58.
11. Liu Y, Chen P-HC, Krause J, et al. How to read articles that use machine learning: Users' Guides to the Medical Literature. JAMA 2019;322:1806-16.
12. Beam AL, Kohane IS. Big data and machine learning in health care. Jama 2018;319:1317-8.
13. Watson DS, Krutzinna J, Bruce IN, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886.
14. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of

diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
15. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115.
16. Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv 2017:171105225.
17. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 2018;378:981.

Cite this article as: Dai L, Yang D, Shen H. The power of clinical data empowered by clinical prediction model: an R tutorial. Ann Transl Med 2020;8(4):77. doi: 10.21037/atm.2020.01.114

