Editorial

Predictive analytics in the era of big data: opportunities and challenges

Zhongheng Zhang

Department of Emergency Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China

Correspondence to: Dr. Zhongheng Zhang, Department of emergency medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China. Email: zh_zhang1984@zju.edu.cn.

Provenance: This is an invited article commissioned by the Editorial Office, *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 Oct 09, 2019. Accepted for publication Oct 22, 2019.

doi: 10.21037/atm.2019.10.97

View this article at: http://dx.doi.org/10.21037/atm.2019.10.97

Big data have changed the way we generate, manage, analyze and leverage data in any industries. There is no exception in clinical medicine where large volume of data is generated from electronic healthcare records, wearable devices and insurance companies (1). This has greatly changed the way we perform clinical studies. Instead of performing data entry and curation manually, the information technology has significantly improved the efficacy of data management. With such a large volume of data, many clinical questions can be addressed by using big data analytics (1-3). Three steps are typically involved in the big data analytics (*Table 1*). The first step is the formulation of clinical questions (4), which can be categorized into three types: (I) epidemiological question on prevalence and incidence and risk factors; (II) effectiveness and/or safety of an intervention; and (III) predictive analytics. The second step is the design of a study, which transforms the clinical question into a study design. For example, the prevalence of catheter-related blood stream infection (CRBSI) as well as its risk factors can be addressed with retrospective or prospective cohort study. A case-control study design can be used to identify risk factors. The effectiveness can be addressed by a randomized controlled trial or an observational study. The third step involves the statistical analysis and/or modelling by using data collected under a certain design.

Among all these big data analytics, the predictive analytics are becoming increasingly important in clinical medicine (5). The use of predictive analytics in clinical medicine includes but not limited to risk stratification, differential diagnosis (classification), prognosis, prediction of disease occurrence and prediction for the effectiveness of a certain intervention (6-8). In other words, the Predictive analytics involve the whole process of the disease course from disease prevention, diagnosis, treatment and finally to the prognosis. For example, from the perspective of disease prevention, smoking is a strong risk factor for the development of lung cancer and thus modifying this factor can help to reduce the risk of lung cancer. If a patient is diagnosed with lung cancer, risk stratification by using genetic and clinical features in a predictive model can help to determine whether surgical intervention and/or chemotherapy should be used. Finally, accurate prediction of the long-term outcome is also important for communication with family members and medical decision making. Thus, the literature involving clinical prediction have witnessed a rapid increase in recent years. Conventionally, predictors are entered into a generalized linear model to estimate a vector of coefficients, and the resulting model can be generalized to samples that are not used for training the model (9). However, the model training process is not straightforward and there is no single approach that can fit for all situations. For example, the generalized linear model is easy to interpret for subject matter audience, but it cannot automatically capture the high-order relationship among covariates (10). In contrast, the sophisticated neural networks and deep learning approaches are capable of modeling any mathematical

Table 1 Examples of big data analytics in clinical medicine highlighting the three steps from clinical question formulation to statistical analytics

Clinical question	Study design	Statistical analytics
What's the prevalence of depression among young children?	Cohort study	Statistical description, multivariate regression model to identify risk factors
Is sodium bicarbonate effective for metabolic acidosis?	Prospective/ retrospective cohort study	Statistical inference; causal inference; and multivariate regression model to adjust for confounders
What is the hospital mortality of patient with severe acute respiratory distress syndrome	Cohort study	Predictive analytics with all kinds of mathematical modeling approaches.

functions, which however is at the cost of interpretability (e.g., these models are considered to be black box algorithm because domain experts cannot easily understand how the predictors/features influence the outcome/label) (11,12).

In a recent special report published in the *Annals of Translational Medicine*, Zhou and colleagues provided a comprehensive tutorial on how to perform predictive modeling (13). There are 16 sections involving variable selection (feature engineering), model calibration, utility, and nomogram for the ease of clinical application. They also discussed some challenging conditions such as the presence of competing risks and the curse of dimensionality. Potential readers of this report include clinical investigators, physicians, and even statisticians. More importantly, the R code for each step of modeling are provided and well explained. For beginners with limited experience in R coding, this can be a good starting point.

However, I want to clarify that the authors have confused the parametric and non-parametric modeling in the first chart. First, let's look at the formal definition for parametric and non-parametric modeling from the textbook *Artificial Intelligence: A Modern Approach* (14). The author stated that:

"*A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model. No matter how much data you throw at a parametric model, it won't change its mind about how many parameters it needs.*"

From this definition, the neural networks should apparently be classified as the parametric modeling approach because there are multiple weights attached to the nodes of the neural network (15). Actually, a neural network with only one layer is simply a linear regression model, and the latter is a prototype of parametric model. The purpose of training a neural network model is to estimate weights and bias for each node, then the weighted sum is passed to the next layer node and there is usually a non-linear activation function to transform the signal. Other machine learning methods such as k-nearest neighbors and decision trees can be safely classified as non-parametric models.

Furthermore, in the prediction evaluation model branch, the authors classified drawing nomogram and building prediction scores into the evaluation process of a model. I have to argue that there is no evaluation of the model at all with these two approaches. The use of risk scores and nomograms are simply the presentation of trained prediction models so that they can be used in clinical practice (16). It has nothing to do with the calibration or discrimination of the model. Nomogram and/or risk scores should be done after the final model is confirmed by using a variety of validation methods. The validations in the training set and external set cannot be considered as conceptually parallel. The external validation should be considered more robust in identifying the problem of overfitting than the internal validation no matter which procedure is used (e.g., there are many statistical methods to perform model validation if there is only one single dataset such as cross validation, simple-split and leave-one-out) (17).

In conclusion, the comprehensive tutorial is timely in the era of big data that it provides practical tools for conducting predictive analytics. With more advanced information technology being applied to patients, a large volume of data can be collected with ease. Thus, the interests in leveraging big data to advance the healthcare are increasing. Predictive analytics is the cornerstone of precision medicine that patients with different clinical characteristics and genetic backgrounds should be treated differently. Although there is a great deal of challenges in leveraging big data to advance the healthcare (18,19), the opportunities are equally abundant.

Acknowledgments

Funding: Z Zhang received funding from Zhejiang Province

Public Welfare Technology Application Research Project (CN) (LGF18H150005) and the National Natural Science Foundation of China (Grant No. 81901929).

Footnotes

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. Wu PY, Cheng CW, Kaddi C, et al. -Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Trans Biomed Eng 2017;64:263-73.
2. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics 2015;8:33.
3. Butte AJ. Big data opens a window onto wellness. Nat Biotechnol 2017;35:720-1.
4. Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. JAMA Intern Med 2014;174:710-8.
5. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Ann Intern Med 2015;162:55-63.
6. Adhikari L, Ozrazgat-Baslanti T, Ruppert M, et al. I Improved predictive models for acute kidney injury with IDEA: Intraoperative Data Embedded Analytics. PLoS One 2019;14:e0214904.
7. Kalagara S, Eltorai AEM, Durand WM, et al. Machine learning modeling for predicting hospital readmission following lumbar laminectomy. J Neurosurg Spine 2018;30:344-52.
8. Wong A, Young AT, Liang AS, et al. Development and Validation of an Electronic Health Record-Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. JAMA Netw Open 2018;1:e181018.
9. Harrell FE. Regression Modeling Strategies. New York, NY: Springer New York, 2001.
10. Zhang Z. Multivariable fractional polynomial method for regression model. Ann Transl Med 2016;4:174.
11. Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20-3.
12. Zhang Z, Beck MW, Winkler DA, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. Ann Transl Med 2018;6:216.
13. 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.
14. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Pearson, 2009.
15. Patel JL, Goyal RK. Applications of artificial neural networks in medical science. Curr Clin Pharmacol 2007;2:217-26.
16. Bonnett LJ, Snell KIE, Collins GS, et al. Guide to presenting clinical prediction models for use in clinical settings. BMJ 2019;365:l737.
17. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140.
18. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016;183:758-64.
19. Gill J, Prasad V. Improving observational studies in the era of big data. Lancet 2018;392:716-7.

Cite this article as: Zhang Z. Predictive analytics in the era of big data: opportunities and challenges. Ann Transl Med 2020;8(4):68. doi: 10.21037/atm.2019.10.97

