Europace (2019) **21**, 1284–1285
doi:10.1093/europace/euz130

POINT OF VIEW

Machine learning in 'big data': handle with care

Zak Loring[1], Suchit Mehrotra[2], and Jonathan P. Piccini[3]*

[1]Duke Clinical Research Institute, Durham, NC, USA; [2]North Carolina State University, Raleigh, NC, USA; and [3]Duke Center for Atrial Fibrillation, Electrophysiology Section, Duke University Medical Center, Durham, NC, USA

Online publish-ahead-of-print 18 May 2019

This commentary refers to 'A novel atrial fibrillation prediction model for Chinese subjects: a nationwide cohort investigation of 682 237 study participants with random forest model', by W.-S. Hu *et al.*, on pages 1307–1312.

Machine learning (ML) has rapidly become an attractive analytic method for tapping into the potential of 'big data' through its ability to uncover novel patterns from complex datasets.[1] The strength of these algorithms lies in their ability to combine predictors in non-linear and highly interactive ways to create predictive models that can outperform traditional linear methods. The ML-based algorithms, however, have had mixed results in risk prediction for cardiovascular outcomes in large population studies.[2,3] Additionally, the complex nature of the ML models makes their clinical interpretation challenging. They are often described as a 'black box' and the opacity of their methods attract scepticism and concern about potential systematic errors in methodology that are difficult to detect.[4]

In this issue of *EP-Europace*, Hu *et al.*[5] used a random forest (RF) model to predict the presence of atrial fibrillation (AF) in a population of more than 680 000 patients. Patients were followed for up to 13 years over which time 2.1% had a diagnosis of AF. The authors included 20 potential International Classification of Diseases, Ninth revision (ICD-9)-based predictors in their model and tested the precision, sensitivity, and discriminatory capacity of their RF model for detection of AF. They report that their model had high precision (F1 value 0.968 and precision value 0.958), high sensitivity (recall value of 0.979), and excellent discrimination (receiver operating characteristic area under the curve 0.948). Said another way, their model would correctly identify presence or absence of AF in 94 out of 100 patients and only misclassify 6 patients. They tested their results in an independent population of 18 million patients and reported similar high performance parameters. The results are striking, but are they plausible?

Random forest models generate predictions by averaging multiple individual decision trees (which have high variance, but have low bias) to create a new model with low variance. The data are randomly sampled (with replacement) to generate individual decision trees that generate classification rules based on the clustering patterns of data within each sample. Each tree generates an outcome prediction for an individual and the modal value of all trees is used for the final classification, similar to 'majority rules' voting. These models do not perform data transformations and thus are more easily understood than more complex ML models such as neural networks.

In the present study, the authors use 10 decision tree estimators that utilized the Gini impurity criterion (a measure of how variance in a parameter may impact prediction) to determine classification rules. They present one of their decision trees in *Figure 1*. The first classifier in this decision tree is the variable 'follow-up time'. 'Follow-up time' is also by far the most important feature listed in *Table 2* with a Gini importance value more than double the next most important feature. In reviewing the methods, we see that patients were followed from 1 January 2000 until they were diagnosed with AF, they withdrew from insurance, or the end of 2013. This means that patients who were diagnosed with AF had systematically shorter follow-up times than those who were not diagnosed with AF solely as an artefact of the study design. Indeed, the mean follow-up in the AF group was 7.1 years; whereas, those without AF had a mean follow-up of 12.7 years, nearly identical to the total study period. Thus, the high performance of the model is expected; not due to its particularly novel handling of the predictive parameters, but rather as a result of inclusion of a parameter directly linked to the outcome of interest by nature of the study design.

This violation of internal validity highlights the danger in reliance on methodology without critical evaluation of both its inputs and outputs. While RF models provide insight into which factors are of highest importance in predicting outcomes, this is a more challenging task when ML algorithms such as neural networks perform non-linear data transformations. One must be mindful of the risks of systematic error when designing studies utilizing these powerful tools.

A first glance at the ML model generated by Hu *et al.* suggests that it performs well in predicting the presence of AF in Chinese populations and significantly outperforms other clinical models of AF. However, closer inspection reveals that these assertions are not supported by their data. The study's methodology dictated its results and in doing so, overshadowed any relationship that may exist between the other included parameters and the outcome of interest. Unfortunately, no clinically relevant conclusions can be drawn regarding the relationship between risk factors and AF based on the

* Corresponding author. Tel: (919) 564-9666; fax: (919) 668-7057. *E-mail address*: jonathan.piccini@duke.edu

presented model and its results. As with all powerful tools, ML algorithms should be used thoughtfully and handled with care.

Funding

National Institute of Health (NIH) T32 training grants (#5T32HL069749 and #T32HL079896, respectively to Z.L. and S.M.).

Conflict of interest: J.P.P. receives grants for clinical research from Abbott, American Heart Association, Boston Scientific, Gilead, and Janssen Pharmaceuticals, and serves as a consultant to Abbott, Allergan, ARCA Biopharma, Biotronik, Boston Scientific, Johnson & Johnson, LivaNova, Medtronic, Milestone, Oliver Wyman Health, Sanofi, Philips, and Up-to-Date.

References

1. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016;**375**:1216–9.
2. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF *et al.* Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol* 2017;**2**:204–9.
3. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li S-X *et al.* Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes* 2016;**9**:629–40.
4. The Lancet Respiratory Medicine. Opening the black box of machine learning. *The Lancet Respir Med* 2018;**6**:801.
5. Hu W-S, Hsieh M-H, Lin C-L. A novel atrial fibrillation prediction model for Chinese subjects: a nationwide cohort investigation of 682 237 study participants with random forest model. *Europace* 2019;**21**:1307–12.

IMAGES IN ELECTROPHYSIOLOGY

doi:10.1093/europace/euz043
Online publish-ahead-of-print 20 March 2019

Bidirectional ventricular tachycardia in ACTH-producing pheochromocytoma

Catarina Quina-Rodrigues[1]*, Joana Alves[2], and Cláudia Matta-Coelho[3]

[1]Department of Cardiology, Hospital de Braga, Braga, Portugal; [2]Department of Emergency Medicine, Hospital de Braga, Braga, Portugal; and [3]Department of Endocrinology, Hospital de Braga, Braga, Portugal

* Corresponding author. Tel: +351 253 027 000; fax: +351 253 027 999. *E-mail address:* catarinaquina@gmail.com

A 70-year-old man presented to the emergency department with syncope. He was normotensive and tachycardic (180 b.p.m.). The 12-lead electrocardiogram showed a bidirectional ventricular tachycardia (BDVT): a narrower QRS with left posterior hemi-block (right axis deviation) and apical exit (negative V1-6) alternating on a beat-to-beat basis with a wider QRS with left anterior hemi-block (left axis deviation) and right bundle branch block. Arterial-blood gas showed severe metabolic alkalosis (pH 7.58) and hypokalaemia (1.48 mmol/L). He was defibrillated twice (ventricular fibrillation) and BDVT resumed after aggressive potassium replacement. Diagnostic work-up revealed adrenocorticotropic hormone (ACTH)- producing pheophomocytoma (APPh) confirmed by histopathology. BDVT is a hallmark of digitalis toxicity and catecholaminergic polymorphic ventricular tachycardia, both excluded. APPh-mediated severe hypokalaemia and increased catecholamines might have acted synergistically towards cardiomyocyte calcium overload, lowering the threshold for delayed afterdepolarizations, triggering this rare arrhythmia.