Machine learning in clinical and epidemiological research: Isn't it time for biostatisticians to work on it?

DOI:10.2427/13245

(Article begins on next page)

Machine learning in clinical and epidemiological research: isn't it time for biostatisticians to work on it?

Machine Learning in Clinical Research Group [(1)]

(1) The Machine Learning in Clinical Research Group - Danila Azzolina (University of Piemonte Orientale), Ileana Baldi (University of Padova), Giulia Barbati (University of Trieste), Paola Berchialla (University of Torino), Daniele Bottigliengo (University of Padova), Andrea Bucci (Marche Polytechnic University), Stefano Calza (University of Brescia), Pasquale Dolce (University of Napoli Federico II), Valeria Edefonti (University of Milan), Andrea Faragalli (Marche Polytechnic University), Giovanni Fiorito (University of Sassari), Ilaria Gandin (Area Science Park, Trieste), Fabiola Giudici (University of Padova), Dario Gregori (University of Padova), Caterina Gregorio (University of Padova), Francesca Ieva (Polytechnic of Milano), Corrado Lanera (University of Padova), Giulia Lorenzoni (University of Padova), Michele Marchioni (University of Chieti-Pescara), Alberto Milanese (University of Rome, La Sapienza), Andrea Ricotti (University of Torino), Veronica Sciannameo (University of Padova), Giuliana Solinas (University of Sassari), Marika Vezzoli (University of Brescia).

CORRESPONDING AUTHOR: *Paola Berchialla, Department of Clinical and Biological Sciences, University of Torino, Torino, Italy*

DOI: 10.2427/13245

Accpted on December 11, 2019

In recent years, there has been a widespread cross-fertilization between Medical Statistics and Machine Learning (ML) techniques (1).

A broad range of ML methods are increasingly being used in many medical fields, such as oncology, internal medicine, cardiology, pediatrics, and genetics (2–4) with a particular focus on the development of prediction tools. For example, in personalized medicine ML techniques have been used to derive the probability of treatment response for each patient (5). In oncology and cardiology, the ML approach has focused on prognosis and risk estimation (6,7). Moreover, ML approaches have been often applied to guide treatment decisions, to counsel patients, and to address the critical steps of clinical trials design (8).

Despite its popularity, it is difficult to find a universally agreed-upon definition for ML. It is widely recognized that the major difference between ML and a traditional statistical approach lies in their purpose. ML methods are focused on making predictions as accurate as possible, whereas statistical models are aimed at inferring relationships between variables. However, many statistical models can make predictions too. On the other hand, ML techniques can provide different degrees of interpretability, from neural networks, which sacrifice interpretability to predictive power, to the highly interpretable lasso regression approach.

Traditional approaches for developing predictive models are mainly regression-based models, such as linear or logistic regression. Such models often use a small number of variables to predict the value of an outcome or the probability of an event and they are ubiquitous in clinical research because they estimate easy-to-interpret parameters (e.g. odds ratios, relative risks, and hazard ratios). However, traditional regression-based models rely on strong assumptions, such as additivity, linearity and distributional assumptions, that may be unrealistic in the clinical practice, where relationships between subject's characteristics and a clinical endpoint are likely to be complex.

ML methods possess several attractive properties, such as flexibility and freedom-from-assumptions, that make them valuable alternatives to traditional statistical approaches in medical research. As identified by Goldstein and colleagues (6), MLs are able to face modeling challenges that are often difficult to address with traditional statistical models. Among them the most important are:

1. Non-linearities. Traditional regression models assume that the effect of a predictor on the endpoint increases (or decreases) uniformly throughout the range of the predictor. Such assumption is not always true in practice, where the relationships between covariates and the outcome may be non-linear. For example, the risk of death is likely to increase sharply with increasing age.

2. Effect modification or statistical interaction. It occurs when the effect of a predictor changes given the values of another variable. For example, it has been observed that air pollution may have a differential effect on adverse cardiovascular events depending on genotypes (9). ML techniques can automatically detect such heterogeneity of effects, which, in contrast, has to be a priori specified putting interaction terms in classical statistical approaches.
3. Few observations and many predictors. Datasets in clinical research are often characterized by a small number of patients/observations and many predictors. In such settings, it is crucial to develop predictive tools able to provide robust estimates. Traditional regression methods are known to have several limitations in such situations, especially when the aim is to select the most relevant risk factors. Although the rise of ML has been associated to an unprecedented wealth of data, several strategies can be adopted to overcome the issues of small dataset in building ML predictive models (10).
4. Multiplicity of models. Many models with different sets of features have nearly the same predictive accuracy. This is partly because in real-world data, it is very common to have some degree of correlation between features. Building a single data model means focusing on only one possible representation of the mapping from features to outcome (11).

In 2001, in its seminal paper "Statistical Modeling: The Two Cultures" (12), Leo Breiman described the data modeling approach and the algorithmic modeling as two contrasting cultures. Over the last two decades, medical statistics and ML have blended more and more. ML techniques hold several potential benefits that are increasing their popularity in clinical research. Their significant spread highlights their crucial role in dealing with the integration of complex biomedical and healthcare data in scenarios where traditional statistical methods show limitations (13).

However, ML models must be appropriately developed, evaluated, and eventually tailored to different situations (14). The growing trend on their application in clinical and epidemiological research requires they are assessed considering the established methodological standards applied for traditional prediction model research. This makes ML techniques another effective and powerful tool for carrying out data analysis. They can be also used instrumentally to traditional statistical methods. The two approaches should be considered as complementary rather than competitive. A proper blending may provide a wide variety of statistical and computational tools for theory testing, knowledge discovery, prediction and decision making. In the end, both ML and medical statistics are concerned with the same question: how do we learn from data?

References

1. Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science, 16(3), 199–231. https://doi.org/10.1214/ss/1009213726
2. Collins, G. S., & Moons, K. G. M. (2019). Reporting of artificial intelligence prediction models. Lancet (London, England), 393(10181), 1577–1579. https://doi.org/10.1016/S0140-6736(19)30037-6
3. Deo, R. C. (2015). Machine Learning in Medicine. Circulation, 132(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593
4. Esteva, A., Kuprel, B., Novoa, R. et al. (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115–118 doi:10.1038/nature21056
5. Forman, G., & Cohen, I. (2004). Learning from Little: Comparison of Classifiers Given Little Training. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004 (Vol. 3202, pp. 161–172). https://doi.org/10.1007/978-3-540-30116-5_17
6. Goldstein, B. A., Navar, A. M., & Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. European Heart Journal, 38(23), 1805–1814. https://doi.org/10.1093/eurheartj/ehw302
7. Harrer, S., Shah, P., Antony, B., & Hu, J. (2019). Artificial Intelligence for Clinical Trial Design. Trends in Pharmacological Sciences, 40(8), 577–591. https://doi.org/10.1016/j.tips.2019.05.005
8. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415
9. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195. https://doi.org/10.1186/s12916-019-1426-2
10. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005
11. Kruppa, J., Ziegler, A., & König, I. R. (2012). Risk estimation and risk prediction using machine-learning methods. Human Genetics, 131(10), 1639–1654. https://doi.org/10.1007/s00439-012-1194-y
12. Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: Are we there yet? Heart (British Cardiac Society), 104(14), 1156–1164. https://doi.org/10.1136/heartjnl-2017-311198
13. Tsagris, M., Lagani, V., & Tsamardinos, I. (2018). Feature selection for high-dimensional temporal data. BMC Bioinformatics, 19(1), 17. https://doi.org/10.1186/s12859-018-2023-7

14. Zanobetti, A., Baccarelli, A., & Schwartz, J. (2011). Gene–Air Pollution Interaction and Cardiovascular Disease: A Review. Progress in Cardiovascular Diseases, 53(5), 344–352. https://doi.org/10.1016/j.pcad.2011.01.001

*