Send Orders for Reprints to reprints@benthamscience.net

EDITORIAL

Machine Learning Accuracy and Big Data in Research on Disease and Health

Electronic health records (EHRs), originated by interactions between patients and the health care system, provide unique medical characteristics of each patient; results from laboratory tests, treatment history, adverse events, and comorbidities. Huge volumes of data are generated from NGS, fMRI, and other sophisticated diagnostic technologies. Such data have different variety; structured, semi-structured or un-structured, and are highly variable; time series, different measurement technologies, … Furthermore, EHRs data can be linked to administrative data, lifestyle data, geographical data, environmental data, social networks data, and more. This rich and heterogeneous data environment is known as *big data* and together with the amazing technological enhancements in *machine learning*, *deep learning*, and *reinforcement learning*, of the last two decades, make it possible, at least in principle, to develop new approaches to design and develop drugs and therapies for almost any disease.

Randomized control trials (RCTs) are commonly recognized as the gold standard for developing and testing drugs and therapies. However, it is well known that RCTs suffer from *external validity*, *i.e.*, they do not take into proper account the difference of patients characteristics between the *target population*, *i.e.*, the population for which we design, develop and test drugs and therapies, and the *study population*, *i.e.*, the population we analyze, by randomized control trial, to develop and test drugs and therapies. On the contrary, EHRs are *observational data*, *i.e.*, data collected under uncontrolled settings, are known to behave better than *interventional data*, *i.e.*, data collected by RCTs, in terms of external validity. However, observational data suffer from *internal validity*, *i.e.*, they could ignore the existence of patients characteristics which are relevant to assess the effectiveness of a drug or therapy, they typically suffer from *confounding bias*.

Machine learning (ML), leveraging on big data, *i.e.*, by combining interventional data and observational data, has the potential to achieve the goal that traditional statistical approaches cannot achieve. In particular, ML methods and algorithms can provide a fundamental contribution to develop and test drugs and therapies which are effective and safe not only at a population level but also, and more importantly at the level of the individual patient, thus by tackling the challenge of *personalized medicine*.

In this special issue we present four papers which make relevant contributions to describe and analyze what has been achieved, at the current state of the art, by applying ML methods and algorithms to big data in research of disease and health. In particular, two papers concern cancer drug design, development and testing, while a third paper is about network inference approaches of the research landscape of microbiota composition studies. The fourth and last paper gives concrete examples of learning-based approaches and learning algorithms in several healthcare fields, including radiology, genetics, electronic health records, and neuroimaging.

The first paper with title "*Deep Hidden Physics Modeling of Cell Signaling Networks*" [1] gives valuable and thought opinions to the subject of the special issue. The authors start by mentioning that according to the WHO, cancer is the second most common cause of death worldwide. Therefore, the authors highlight how the social and economic damage caused by cancer is high and continuously rising. The paper mentions that in Europe, the annual direct medical expenses alone amount to more than €129 billion, and thus there is an urgency to design, develop and test new and sustainable therapeutics. Indeed, such a relevant and urgent need is currently not met by the pharmaceutical industry, because only 3.4% of cancer drugs entering Phase I clinical trials get to market. The authors recognize phosphorylation sites as parts of the core machinery of kinase signaling networks, which are known to be dysfunctional in all types of cancer. Indeed, kinases are the second most common drug target, yet, because these inhibitors block all functions of a protein they commonly lead to resistance development and increased toxicity. Furthermore, the authors explain that to facilitate global and mechanistic modeling of cancer and clinically relevant cell signaling networks, the community will have to develop sophisticated data-driven deep-learning and mechanistic computational models that generate *in silico* probabilistic predictions of molecular signaling network rearrangements causally implicated in cancer.

In the second paper with title "*Big Data to Knowledge: Application of Machine Learning to Predictive Modeling of Therapeutic Response in Cancer*" [2], the authors present and discuss how the availability of high throughput technologies, establishment of large molecular patient data repositories, and advancement in computing power and storage, allowed to elucidate complex mechanisms implicated in therapeutic response in cancer patients. The paper states that breadth and depth of the available data, alongside with experimental noise and missing values, require a sophisticated human-machine interaction that would allow effective learning from complex data and accurate forecasting of future outcomes, ideally embedded in the core of machine learning design. The authors present and discuss machine learning techniques utilized for modeling of treatment response

1389-2029/21

in cancer. In particular, they introduce and describe non-parametric models as random forests, and artificial neural networks, as well as parametric models as support vector machines, linear and logistic regression. Mathematical foundations, strengths and weaknesses of the presented ML models and algorithms are discussed together with alternative approaches in light of their application to therapeutic response modeling in cancer. The paper closes the rich and valuable analysis by hypothesizing that the increase in the number of available patient profiles, together with the possibility to temporally monitor patient data, will allow to define even more complex techniques, such as deep learning and causal analysis, as central players in therapeutic response modeling.

The third paper, with title "*Modeling Microbial Community Networks: Methods and Tools*" [3], makes a relevant contribution by presenting and discussing the current research landscape about microbiota composition studies. Indeed, such studies are of extreme interest, since it has been widely shown that resident microorganisms affect and shape the ecological niche they inhabit. According to authors this complex micro-world is characterized by different types of interactions. Therefore, understanding these relationships provides a useful tool for decoding the causes and effects of communities organizations. In this respect, the paper clarifies that Next-Generation Sequencing technologies allow to reconstruct the internal composition of the whole microbial community present in a sample. The paper continues by showing how sequencing data can be investigated through statistical and computational methods coming from network theory to infer the network of interactions among microbial species. The authors point out that since there are several network inference approaches in the literature, their paper tries to shed light on their main characteristics and challenges. Therefore, the paper provides a useful tool not only to researchers interested in applying the methods, but also to researchers who want to develop new methods and algorithms. The paper also discusses frameworks used to produce synthetic data, starting from the simulation of network structures up to their integration with abundance models, with the aim of clarifying the key points of the entire generative process.

The fourth and last paper of the special issue has the following title "*Machine Learning in Healthcare*" [4] and describes recent advancements in artificial intelligence and machine learning technology. The paper also discusses how such advancements brought on substantial strides in predicting and identifying health emergencies, disease populations, and in predicting disease state and immune response. The paper mentions that skepticism remains regarding the practical application and interpretation of results from ML-based approaches in healthcare settings, even if the adoption of these approaches is expected to increase at a rapid pace. The paper provides a brief overview of machine learning-based approaches and learning algorithms including supervised, unsupervised and reinforcement learning along with practical examples. A second relevant contribution of the paper consists in the discussion of different applications of ML with specific reference to radiology, genetics, to the use of EHRs, and to neuroimaging. The paper closes by discussing different impacts of ML adoption in disease and health. In particular, the authors discuss risks and challenges of ML application to healthcare, with specific reference to privacy and ethical concerns by providing suggestions for future applications.

In summary, ML offers an unprecedent and extraordinary opportunity for exploiting big data, combining interventional and observational data, to design, develop and test individualized drugs and therapies. However, it is extremely important to be aware of conditions under which the developed ML models will be reliable, understandable and accountable, as well as to take into proper account ethical concerns of automated patient treatment. I personally think this special issue makes several relevant contributions in the direction to present and clarify different issues of ML and big data for disease and health. Furthermore, the special issue is a starting point for discussion on such a relevant aspect of learning and automated treatment design and development in health care and medicine.

I would like to thank all authors contributing the special issue, for providing high quality papers, the Editor in Chief, Prof. Christian Neri, and all member of the editorial staff of the Journal of Current Genomics for being that helpful in developing this special issue.

REFERENCES

[1] Seeger, M.; Longden, J.; Klipp, E.; Linding, R. Deep hidden physics modeling of cell signaling networks. *Curr. Genomics*, **2021**, *22*(4), 239-243.

[2] Panja, S.; Rahem, S.; Chu, C.J.; Mitrofanova, A. Big data to knowledge: Application of machine learning to predictive modeling of therapeutic response in cancer. *Curr. Genomics*, **2021**, *22*(4), 244-266.

[3] Cappellato, M.; Baruzzo, G.; Patuzzi, I.; Camillo, B.D. Modeling microbial community networks: methods and tools. *Curr. Genomics*, **2021**, *22*(4), 267-290.

[4] Habehh, H.; Gohel, S. Machine learning in healthcare. *Curr. Genomics*, **2021**, *22*(4), 291-300.

Fabio Stella
(Guest Editor)
Department of Informatics, Systems and Communication
University of Milan-Bicocca, 20126
Milan, Italy
Email: fabio.stella@unimib.it