Risk prediction: Methods, Challenges, and Opportunities

Ruowang Li
Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, California, USA
Email: ruowang.li@cshs.org

Rui Duan
Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
Email: rduan@hsph.harvard.edu

Lifang He
Department of Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania, USA
Email: lih319@lehigh.edu

Jason H. Moore
Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, California, USA
Email: jason.moore@csmc.edu

The primary efforts of disease and epidemiological research can be divided into two areas: identifying the causal mechanisms and utilizing important variables for risk prediction. The latter is generally perceived as a more obtainable goal due to the vast number of readily available tools and the faster pace of obtaining results. However, the lower barrier of entry in risk prediction means that it is easy to make predictions, yet it is incredibility more difficult to make sound predictions. As an ever-growing amount of data is being generated, developing risk prediction models and turning them into clinically actionable findings is crucial as the next step. However, there are still sizable gaps before risk prediction models can be implemented clinically. While clinicians are eager to embrace new ways to improve patients' care, they are overwhelmed by a plethora of prediction methods. Thus, the next generation of prediction models will need to shift from making simple predictions towards interpretable, equitable, explainable and ultimately, casual predictions.

Keywords: Risk Prediction; Methodology; AutoML, Explainable Artificial Intelligence, Federated Learning, Model Interpretation.

1. Introduction

The purpose of this workshop is to introduce and discuss the current and future of risk prediction in the context of disease and epidemiological research. We will discuss the pressing topics ranging from data sources to model implementation. Our speakers will discuss the most commonly used data sources, e.g., genetics, imaging, clinical, and epidemiological data, for developing the

prediction models. A number of novel risk prediction methods, including automatic machine learning (AutoML), explainable artificial intelligence (XAI), and polygenic risk score, will be presented. Issues regarding how to handle the high dimensionality of the features will be discussed from the perspective of accuracy and computational scalability. Data privacy considerations during the construction and dissemination of prediction models will be addressed. Furthermore, model-based and post-hoc analysis of prediction models, including the biases and uncertainty quantification, model interpretation, and fairness and diversity of the prediction results, transferability and generalizability of the models to different populations and datasets will be thoroughly discussed. Finally, the current progress and future perspective regarding the validation and clinical implementation of the risk prediction models will be reviewed.

2. Machine learning

Recent advances in machine learning (ML) methods, combined with the rapidly increasing availability of healthcare data, forebode an avalanche of explorations of ML in medical research. Since risk prediction tasks constitute a large portion of the applications of ML in medicine, knowledge on how to develop, implement and evaluate risk prediction models, as well as interpret the results on their basis is critical for enhancing the model quality, transparency, trust and for decreasing the instances of bias. This workshop provides a roadmap to help refine and enhance understanding of risk prediction and assessment by focusing on all stages of developing and validating risk prediction models.

2.1 Automatic Machine Learning

One of the many challenges of machine learning is the selection of the method to use and the tuning of its hyperparameters. This is a challenge for both experts and beginners because there are dozens of methods and each looks at the data in a different way. It is difficult to know which method is most appropriate when using machine learning to develop risk models. Automated machine learning (AutoML) seeks to address this issue by exploring a wide range of models and hyperparameters with minimal user input. Maduchi et al. (2022) recently reviewed automated machine learning for the genetic analysis of complex traits. One of these methods, the Tree-Base Pipeline Optimization Tool (TPOT), has been applied to genomics data (Le et al. 2020) and uses expression trees to represent machine learning pipelines with operators including feature selectors, feature transformers, feature engineering algorithms, and a wide range of machine learning algorithms all available from the sci-kit learning library. Pipelines are explored and optimized using genetic programming with multi-objective optimization and cross-validation to limit overfitting. Manduchi et al. (2022) demonstrate the application of TPOT to the genetic analysis of coronary artery disease (CAD) using genome-wide association study (GWAS) data from UK Biobank. A central focus of this study was prioritizing genes based on their druggability and pharmacologic relevance to CAD. The TPOT algorithm was able to automatically identify an optimal machine learning pipeline for predicting CAD with evidence of genetic heterogeneity revealed by feature importance score methods. This study is used as an example to demonstrate the potential for AuoML to inform the development of genetic risk models for common disease.

3. Statistical modeling

Statistical modeling plays an important role in risk prediction, which has a broad application in clinical science, epidemiology, and health services. With the growing availability and variety of real-world healthcare data sources, such as claims data and electronic health records, there are emerging statistical challenges that need to be addressed for constructing more reliable and generalizable risk prediction tools. In this workshop, we discuss advanced statistical methods that address the following challenges (1) prediction models with limited and imperfect labels (2) building risk prediction models for underrepresented populations with limited data (3) combining data from multiple sources to improve the generalizability and transferability of risk prediction models. In addition to the methods, we will also discuss the theoretical insights and examples of potential real-world applications.

4. Conclusion

Our workshop puts an even focus on all stages of developing and validating risk prediction models. Rather than focusing exclusively on the methodologies, we believe by structuring a more balanced workshop theme, the speakers and the audiences will have more opportunities to exchange ideas and viewpoints. Discussion sessions would also be employed to break up the talks and to provide a venue for general dialog around themes that have evolved from the lectures.

References

1. Manduchi E, Romano JD, Moore JH. The promise of automated machine learning for the genetic analysis of complex traits. Hum Genet. 2022 Sep;141(9):1529-1544.

2. Manduchi E, Le TT, Fu W, Moore JH. Genetic Analysis of Coronary Artery Disease Using Tree-Based Automated Machine Learning Informed By Biology-Based Feature Selection. IEEE/ACM Trans Comput Biol Bioinform. 2022 May-Jun;19(3):1379-1386.

3. Le TT, Fu W, Moore JH. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics. 2020 Jan 1;36(1):250-256.