Abstract: Scientific and technological advances in measurement and sensor devices, databases, and storage systems have made it possible to efficiently collect, store and retrieve huge amounts of and different kinds of data. However, when it comes to the analysis of such data, we have to admit that our ability to generate big data has far outstripped our analytical ability to make sense of it. This is true in practically all fields, and the field of medicine and healthcare is no exception to it, where data science techniques are increasingly playing a major role in complementing physician knowledge and aid in clinical decision support. The need for such data-driven analytics in healthcare is also emphasized by initiatives from the National Institutes of Health (NIH), such as Big Data to Knowledge (BD2K). In this talk, I would present a typical workflow for such predictive analytics involves data collection, data transformation, predictive modeling, evaluation, and deployment, with each step tailored to the end goals of the project. To illustrate each of these steps, we shall take the example of recent advances in such predictive analytics on lung cancer electronic healthcare records (EHR) data from the Surveillance, Epidemiology, and End Results (SEER) program as well as older adults EHR data from Northwestern Memorial Hospital. The resulting best performing models have been deployed as both generic and disease-specific healthcare outcome calculators.

References:


Funding:

NSF awards IIS-1343639, CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330; and Northwestern Data Science Initiative