Personalized Risk Predictions with Deep Learning Methods in the Presence of Missing and Biased Electronic Health Record and Genomics Data

Hua Zhong

NYU Langone Health
E-mail: judy.zhong@nyumc.org

Abstract: Currently available risk prediction methods are limited in their ability to deal with complex, heterogeneous, and longitudinal data such as that available in electronic health records (EHRs) and genomics studies. Recurrent neural networks (RNNs) have shown significant promise in this context, where (essentially) the model learns a latent representation for a patient state over time, updating the state when new covariate measurements arrive, in a flexible non-linear manner. These models have been shown to be very effective in modeling signals such as speech and text sequences, leading to prediction models that are significantly more accurate than previous non-RNN approaches. However, it is not straightforward how to directly apply a conventional RNN to EHR data involving(a) a significant amount of missing data (many covariates such as lab tests might not be measured during a particular patient visit), and (b) asynchronous measurements(patients show up at varying time-intervals). We propose a patient-specific RNN to learn the time-to-event distributions through flexibly incorporating both missing and asynchronous measurements over time. We demonstrate the efficacy of our approach by applying it to a real-world longitudinal EHR dataset to predict cardiovascular disease (CVD) in patients with type 2 diabetes (T2DM).