Learning from EMR/EHR data to estimate treatment effects using high dimensional claims codes

Ronghui Xu

University of California
E-mail: rxu@ucsd.edu

Abstract: Our work was motivated by the analysis projects using the linked US SEER-Medicare database to studying treatment effects in men of age 65 years or older who were diagnosed with prostate cancer. Such data sets contain up to 100,000 human subjects and over 20,000 claim codes. The data were obviously not randomized with regard to the treatment of interest, for example, radical prostatectomy versus conservative treatment. Informed by previous instrumental variable (IV) analysis, we know that confounding mostly likely exists beyond the commonly captured clinical variables in the database, and meanwhile the high dimensional claims codes have been shown to contain rich information about the patients' survival. Hence we aim to incorporate the high dimensional claims codes into the estimation of the treatment effect. The orthogonal score method is one that can be used for treatment effect estimation and inference assuming only consistency under the high dimensional hazards outcome model and the high dimensional conditional treatment model. In addition, we show that further refinement of the approach has doubly-robust properties in high dimensions: the resulting estimator is consistent when either of the hazards model or the treatment model is misspecified, as long as the other model is correct. We also develop a novel sparsity doubly robust result, where either the outcome or the treatment model can be a fully dense high-dimensional model.