Learning survival from electronic medical/health records (EMR/EHR) data using high dimensional claims codes

Abstract:

Our work was motivated by the analysis projects using the linked US SEER-Medicare database to study mortality 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. For studying mortality the number of deaths are the “effective” sample size, so here we are in the situation of p is greater than n which is referred to as having high-dimensional predictors. In addition, a patient might die of cancer, or of other causes such as heart disease etc. These are referred to as competing risks. How to best perform prediction which inevitably involves variable selection for this type of complex survival data had not been previously investigated. Interest may also lie in comparing treatments such as radical prostatectomy versus conservative treatment. In this case the data were obviously not randomized with regard to the treatment assignments, and confounding most likely exists, possibly even beyond the commonly captured clinical variables in the SEER database. We will showcase research work done by our former PhD students from the UCSD Math Dept to account for such unobserved confounding, as well as efforts to make use of the high dimensional claims codes which have been shown to contain rich information about the patients survival.

Special Note:
website: https://sites.google.com/view/awmcolloquiumatucsd/home

Hosts: AWM Organizers Kristin DeVlemining and Ruth Luo

Thursday, February 4, 2021
4:00 PM
https://ucsd.zoom.us/j/94080232559