Mathematics, Statistics, and Computer Science **@ UIC**

Special Colloquium

Leveraging Digital Data for Clinical Research

Jessica Gronsbell (Alphabet's Verily Life Sciences)

Abstract: The widespread adoption of electronic health records (EHR) and their subsequent linkage to specimen biorepositories has generated massive amounts of routinely collected medical data for use in translational research. These integrated data sets enable real-world predictive modeling of disease risk and progression. However, data heterogeneity and quality issues impose unique analytical challenges on the development of EHR-based prediction models. For example, ascertainment of validated outcome information, such as the presence of a disease or treatment response, is particularly challenging because it requires manual chart review. Outcome information is therefore only available for a small subset of patients in the cohort of interest, unlike the traditional setting where this information is available for all patients. In this talk, I will discuss semi-supervised and weakly-supervised learning methods for predictive modeling in such constrained settings where the proportion of labeled data is very small. I demonstrate that leveraging unlabeled examples can improve the efficiency of model estimation and evaluation procedures, which in turn substantially reduces the amount of labeled data required for developing prediction models.

Friday, January 24 at 3:00 PM in 636 SEO