Estimation and Analysis of EHR-derived Phenotypes

Dr. Rebecca Hubbard
University of Pennsylvania
School of Medicine, Departments of Biostatistics, Epidemiology and Informatics

Abstract:
A first step in almost all analyses conducted using EHR data is characterization of patient phenotypes. However, EHR-based phenotyping is hampered by complex missing data patterns and heterogeneity across patients and healthcare systems in the amount and type of data that is available. As a result, not only are EHR-derived phenotypes expected to be imperfect, but they will very often feature exposure-dependent differential misclassification, which can bias analyses towards or away from the null. In this talk I will first review approaches to EHR-based phenotyping, highlighting how missing data affect phenotype estimation. I will then review some results on the implications of using EHR-derived phenotypes with differential misclassification for bias and type I error of subsequent association studies using these phenotypes as outcomes. Finally, I will present an approach to correcting for phenotyping error that does not require knowledge of sensitivity and specificity of the phenotype. The overall goal of this presentation is to improve awareness of phenotyping error, its implications for analyses, and available options for valid analysis of EHR-derived phenotypes.