Efficient Use of EMR for Discovery Research

A. Chakrabortty\textsuperscript{1}, J. Gronsbell\textsuperscript{1} and T. Cai\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1} Dept of Biostatistics, Harvard University
abhi060988@gmail.com; jlg735@mail.harvard.edu; tcai@hsph.harvard.edu
*Presenting author

\textbf{Keywords.} Classification; Electronic Medical Records Data; Model mis-specification; Robustness; Semi-supervised learning

In clinical practice, patients with the same disease diagnosis often differ in outcomes and response to treatment. The ability to both classify and predict disease phenotypes would be a valuable asset in clinical decision-making. Large datasets containing both a wealth of clinical and experimental data now exist as a result of the increasing adoption of electronic medical records (EMR) linked with specimen bio-repositories. These datasets allow for data driven classification and prediction of sub-phenotypes and investigation of shared risk factors across a group of phenotypes. In this talk, I’ll discuss various statistical and informatics methods that illustrate both the challenges and potential opportunities that arise from analyzing EMR data. For example, obtaining validated phenotype information is a major bottleneck in EMR research, as it requires laborious medical record review. Thus gold standard labels are typically available only in a small training set nested in a large cohort. In contrast, data on the clinical predictors of the phenotype are often available on all subjects. To improve phenotype definition, we developed robust semi-supervised learning methods that can leverage such rich source of auxiliary information. These methods are illustrated with an EMR cohort of RA patients.