Abstract: Selection of important variables contributing to the “true model” generating the data is one of the most challenging tasks in biomedical, social, and public health science. The problem gets complicated even further in the presence of missing data. A common and simplistic practice is Complete Case Analysis (CCA), which involves list-wise deletion of missing observations. The resulting analysis is highly restrictive, inefficient and often results in a biased model unless the missing pattern is completely at random (MCAR). Multiple imputation is a popular approach to tackle the problem of data analysis with missingness, under the missing at random (MAR) condition. However, combining the results of variables selected via Multiple Imputed (MI) datasets is nontrivial. This is because conventional method of combining parameter estimates obtained via MI datasets using Rubin's rule is not applicable for variable selection. In this talk we present two intuitive approaches for doing variable selection under MAR assumption for linear model. Our approaches are closely related to penalized variable selection methods (e.g. Lasso, SCAD, MCP etc.) but requires significant further innovation due to missing data. The performance of the proposed methods is evaluated using extensive simulation studies and an application on Mental Health studies related to depression.

Lunch will be provided, please RSVP to Tammy Falla at (tammy.falla@umassmed.edu)