GENETIC ASSOCIATION TESTING WITH IMPERFECT PHENOTYPES DERIVED FROM ELECTRONIC HEALTH RECORDS

JENNIFER A. SINNOTT
Department of Statistics, Ohio State University, Columbus, OH, USA
e-mail: jsinnott@stat.osu.edu

Electronic health records linked to blood samples form a powerful new data resource that can provide much larger, more diverse samples for testing associations between genetic markers and disease. However, algorithms for estimating certain phenotypes, especially those that are complex and/or difficult to diagnose, produce outcomes subject to measurement error. Work is needed to determine best practices for analyzing such data. To this end, we recently proposed a method for analyzing case-control studies when disease status is estimated by a phenotyping algorithm; our method improves power and eliminates bias when compared to the standard approach of dichotomizing the algorithm prediction and analyzing the data as though case-control status were known perfectly. The method relies on knowing certain qualities of the algorithm, such as its sensitivity, specificity, and positive predictive value, but in practice these may not be known if no “gold-standard” phenotypes are known in the population. A common setting where that occurs is in phenome-wide association studies (PheWASs), in which a wide range of phenotypes are of interest, and all that is available for each phenotype is a surrogate measure, such as the number of billing codes for that disease. We propose a new method to perform genetic association tests in this setting, which improves power over existing methods that typically identify cases based on thresholding the number of billing codes. In this talk, I will describe these methods, and present applications to studies of rheumatoid arthritis in the Partners Healthcare System.