The availability of vast databases of health information derived from electronic health records (EHR) have generated enormous enthusiasm for conducting EHR-based research. Analyses of EHR data are potentially more timely and efficient than clinical trials and other studies requiring primary data collection. However, there are important limitations to EHR data including lack of gold-standard outcomes for many conditions and heterogeneity in data quality and availability across patients. In this presentation, we consider both the potential and challenges for conducting research using EHR data, focusing on the motivating example of studying early-life risk factors for pediatric type 2 diabetes. Past studies investigating early-life risk factors for type 2 diabetes have used longitudinal cohort data collected over the course of many years. In contrast, EHR data from routine clinical care can provide longitudinal information on early-life risk factors as well as subsequent health outcomes without requiring any additional data collection. In this study, we introduce a Bayesian joint phenotyping and BMI trajectory model to address data quality challenges in an EHR-based study of early-life BMI and type 2 diabetes in adolescence. We demonstrate that EHR data coupled with modern methodologic approaches can improve efficiency and timeliness of studies of childhood exposures and rare health outcomes. The overarching objective of this presentation is to raise awareness of the opportunities and pitfalls associated with EHR-based research and introduce statistical tools to mitigate some of the challenges.

This talk is suitable for undergraduate students.