Electronic Health Record Phenotyping using Anchor-Positive and Unlabeled Patients

Jinbo Chen

University of Pennsylvania
E-mail: jinboche@pennmedicine.upenn.edu

Abstract: Phenotyping patients in electronic health records (EHRs) conventionally relied on algorithms learned from labeled cases and controls. Assigning labels requires manual medical chart review and therefore is labor intensive. We developed a phenotyping method when identification of gold-standard controls is prohibitive so that a validation set is not available. Our method relies on a random subset of cases, which can be specified using an expert-derived anchor variable that has excellent positive predictive value and sensitivity independent of predictors. Adopting a maximum likelihood approach to efficiently leveraging data from the anchor-labeled cases and unlabeled patients to develop logistic regression phenotyping models, we propose novel statistical methods for internally assessing model calibration and predictive performance measures. Upon identification of an anchor variable by clinical experts that is scalable and transferable to different practices, our approach should facilitate development of scalable, transferable, and practice-specific phenotyping models. Through phenotyping two cardiovascular conditions in Penn Medicine EHRs, we demonstrate that our method enables accurate semi-automated EHR phenotyping with minimal manual labeling and therefore is expected to greatly facilitate EHR clinical decision support and research.