Making logitudinal data less complex

JH Creed, G Aden-Buie, K Yamoah, TA Gerke

**Background:**

Longitudinal data, data where information about exposures, outcomes and other variables are collected at multiple time points, is becoming increasingly common in epidemiological and clinical research, due in part to the increasing popularity of integrated data services such as Avatar and flatiron.

Challenges of utilizing longitudinal data, and in particular longitudinal data from electronic health records (EHR), include developing proper methods for working with repeated measures and time-varying covariates. With repeated measures, researchers need to address intra-participant correlation between measures, while time-varying covariates require unique causal methods to avoid reverse causation. These challenges become increasingly more difficult when dealing with irregular time intervals.

**Methods:**

An R pipeline for cleaning and organizinging complex longitudinal data from EHRs that outputs easily reproducible analytic datasets. These datasets are largely organized around two main datasets: one for static variables with one row per patient and one for time-varying covariates, with muitple rows per patient with each row indicating a unique time for each patient. Our pipeline utalizes tidy principles, for their readability, and the drake workflow for it’s efficiency handling changes to the pipeline.

**Results:**

Utilizing the prostate cancer data from flatiron, we used this pipeline to examine trends in time-to-metastasis and castration-resistant prostate cancer (CRPC) by race. Men who self-idenitified as Black or African American (AAM) developed mets and CRPC on average a year sooner than self-identified White men (EAM) (4.07 vs 5.03 years; mets and 5.20 vs 6.14 years; crpc). For those who we could calculate time to mets, 51% of AAM presented with de novo mets, while only 43% of EAM did. Even when excluding those with de novo mets and crpc, AAM experience shorter time to mets by an average of 7 months and shorter time to CRPC by 11 months.

**Conclusions:**

We propose a novel R-based pipeline that allows for easier application of longitudinal EHR data and may accelerate the integration of such data into ongoing cancer research.