Utility of Family Data Extracted from Hospital Records

Susmitha Bommini
Michigan Technological University

Susmitha Bommini, Scott Hebbring, PhD, and Brooke Delgoffe, MS, *Center for Precision Medicine Research and Office of Research Computing Analytics*

Background: Family histories serve as a valuable resource for identifying and managing disease risks in genetic research. Traditional methods of obtaining information on family histories, however, are cumbersome. Previous studies revealed the use of electronic health records (EHR) data to generate E–pedigrees using the family mapping algorithm (FMA). While EHRs offer standardized data collected through pre-defined formats, they often lack historical depth in time periods prior to EHR implementation. To help fill such historical gaps, this project focused on utilization of hospital paper records. Although paper records contain a wealth of longitudinal, family health data that predated the EHR, they present significant challenges for utilization. The purpose of this project is to create actionable family history datasets from paper records for use in the FMA.

Methods: Paper records were reviewed by staff to create the electronic data used in this project. Personal identifiers (names, addresses, and contact information) were extracted along with other temporal information and placed in electronic forms in a REDCap database. Then we conducted exploratory data analysis. Due to non-standard formats, these paper records needed extensive pre-processing to make the data trustworthy. Our data included information from birth and non-birth (primary patient) records. A single paper chart can define multiple sets of identifiers: baby, mother, and father (for birth records), patients being seen (adult/child), and their multiple emergency contacts. We re-organized the data to treat sets of identifiers listed in different capacities in a record as a unique person longitudinally and populate necessary fields from corresponding sources. We created unique identifiers for each individual to ensure consistency across different input files. Given the limited availability and missingness of time variables, we employed logical conditions to construct timeframes across the information fields. Additionally, extensive data cleaning was performed to standardize each variable due to high prevalence of manual data entry errors. We used various data pre-processing and mapping techniques to de-identify the data and produce final structured datasets.

Results: With the utilization of SAS, Python, and MS Excel, we transformed unstructured paper records data into an actionable format for use in predicting family members. From around 67k records, we identified 100k+ individuals. Input files (names, demographics, addresses, emergency contacts, and patient information) were created in compliance to FMA requirements.

Conclusions: This approach of integration of paper records data with an EHR underscores the effective utilization of family data in genetic research. Further steps in this extensive project include additional data cleaning and transformation techniques to enhance the quality and accuracy of input data and improve the algorithm's ability to construct E-pedigrees.