Identifying Families in Electronic Health Records

Akash Choudhuri
University of Iowa

Akash Choudhuri, Scott Hebbring Ph.D., & Brooke Delgoffe, M.S.
Center for Precision Medicine Research & Office of Research Computing and Analytics

Background: The widespread adoption of Electronic Health Records (EHR), along with other technological advancements, has simplified the storage and querying of multimodal patient data. This has helped converge genomics and other digital health data, where the ability to reconstruct familial relationships using EHR data has become possible with broad applications (e.g., early detection of hereditary diseases, personalized care). Family mapping algorithms (FMA) enable large-scale pedigree reconstruction using data routinely collected as part of healthcare delivery. Such algorithms are in their infancy though and require additional documentation and logic refinement.

Methods: The existing E-Pedigrees FMA codebase was reviewed and documentation of the algorithm added to facilitate a richer understanding of the order of operations. The documentation process assisted in resolving gender mismatch issues wherein, for example, FMA was predicting 734 males as mothers and 892 females as fathers. Resolution was achieved via modification of the algorithm to map emergency contacts to medical history numbers (MHNs) with name and date of birth and finally query for the MHN in the algorithm (instead of fuzzy matching on name and address). The E-pedigrees pipeline was also supplemented with additional input files for known relationships. These files integrate ground truth (highest confidence) and supplementary (low confidence) data, allowing the pipeline to consider varying degrees of confidence for named relationships previously provided only as emergency contacts. Finally, a new feature-based technique was formulated to aid machine learning models to automatically derive data-dependent decisions to assign individuals to families.

Results: E-Pedigrees software was supplemented with documentation and commentary highlighting the overall logic and exact decision points. Preliminary results of fixing known issues led to 828 new mappings between patients and families, with the discovery of 1,147 new families. Additional validation with the PMRP database corresponded to an accuracy of 96.5%, a 1.9% increase from the previous version of the algorithm. The addition of supplementary ground truth data in the form of PMRP (most confident) and obituaries (least confident) led to 1,092,838 MHNs being mapped to 289,753 families, an increase of 47,141 MHNs and 18,834 families over the previous version of the algorithm.

Conclusions: The incorporation of documentation sets the foundation of explainability of the E-pedigrees software and will aid in future developments. Additionally, the software updates now can ingest ground truth data of varying degrees of confidence, which provides the platform for broader applications of this feature-based technique. These refinements will help optimize automated pedigree analyses and translating genetic insights into more effective clinical interventions.