

An Algorithmic Pipeline for Inferring Family Relationships from Collections Data



Anita Michael
University of Pittsburgh

Anita Michael, Scott Hebbring, PhD and Brooke Delgoffe, MS.
Center for Precision Medicine Research & Office of Research Computing and Analytics

Background: Family history is essential to identify disease patterns and guide clinical management. Existing electronic health record (EHR) based pedigrees (e-pedigrees), generated using Family Mapping Algorithms (FMA), organize family data, but may miss relationships not documented in routine care. Collections data from Alliance Collection Agencies, Inc. include roles such as debtors, responsible parties, spouses and comakers that may provide indirect evidence of family relationships. In this project, we developed a reproducible pipeline

to transform data from collections agencies into a family history dataset for use in validating and enriching FMA predicted family structures.

Methods: General accounts, spouse, comaker, extended patient field and previous address files were deduplicated, standardized and restricted to non-singleton family evidence. Person-level records were created using names, dates of birth (DOB), dates of death (DOD), social security numbers (SSN), medical health numbers (MHN), addresses, client account numbers (CAN) and relationship roles. Exact and fuzzy graph links based on shared CANs, MHNs, SSNs, name-DOB and name-CAN keys, and address/name similarity were used to deduplicate people and recursively merge family components. Records were enriched with EHR data through hierarchical matching using existing MHNs, year of birth and unique name-based matches. Relationship inference was performed within final non-singleton families using spouse as the family anchor. Spouse anchors were selected from direct spouse fields, unique or switched debtor-responsibility party patterns, and single debtor-responsible party pairs supported by surname repetition. Anchored and unanchored family members were classified as spouse, parent, child, sibling, grandparent, grandchild, spouse-likely or spouse/sibling using role labels, birth-year thresholds, gender and surname evidence. Inferred relationship context within families and pairwise age rules were applied when no anchor was available. Collections derived relationship pairs were then compared with FMA by MHN and relationship class.

Results: The pipeline identified 83,110 non-singleton collections-derived families containing 262,432 individuals, of whom 182,679 (69.6%) had an MHN. Relationship inference produced 218,595 person-pairs; 149,267 (68.3%) had MHNs for both individuals. Among 89,031 pairs present in both collections and FMA, 75,067 (84.3%) had concordant relationship classes, and collections contributed 56,614 comparable pairs not identified by FMA. Across all within-family collections pairs, 209,273 of 425,928 (49.1%) had MHNs for both individuals; among these, 144,423 (69.0%) mapped to the same FMA family, 24,181 (11.6%) mapped to different FMA families, and 40,669 (19.4%) had at least one MHN absent from FMA. Manual PPV was 83.34% for collections-derived relationships and 90% for FMA relationships.

Conclusions: Collections data has potential to provide complementary familial relationship evidence for validating and expanding FMA derived e-pedigrees. Further validation against known family structures is essential to assess whether they improve pedigree completeness and accuracy.