

International Journal of Population Data Science

Journal Website: www.ijpds.org



Identifying Family Relationships from Electronic Health Records: A Machine Learning Approach

Toan Ong¹, Abhinav Pundir¹, Andrew Hill¹, Michael Kahn¹, Shaun Grannis², Titus Schleyer², Sarah Kautz¹, Daniel Lindberg¹, Bethany Kawn¹, and Lisa Schilling¹

¹University of Colorado Anschutz Medical Campus, Aurora, USA

²Regenstrief Institute, Indianapolis, USA

Background

Identifying family relationships within electronic health records (EHR) could enable rich environmental and genetic studies, but traditional record linkage methods focus on identifying duplicate records of individual patients and cannot detect familial connections such as parent-child, sibling, and partner relationships using demographic variables alone.

Objective

To develop and evaluate random forest models for identifying five family relationship types using only EHR demographic identifiers.

Methods

Gold-standard relationships (mother-child, father-child, sibling, twin, partner) were derived from birth records in the Indiana Natality Dataset linked to the Indiana Network for Patient Care (>18 million patients). Models utilized 14 demographic features, including age differences, name edit distances, address similarity, and sex encoding. Hyperparameter optimization employed random search with 5-fold cross-validation across 100 configurations per relationship. Validation included standard 70/30 splits, 20-year temporal shift testing, and manual expert review of 100 predictions per relationship type.

Results

Models achieved F1-scores of 0.93-0.99 across all relationship types, with twin identification performing highest (precision/recall/F1: 0.99). True positive predictions concentrated heavily in high-confidence ranges (0.9-1.0 probability: 81-94% across models). False positives predominantly involved single matching identifiers with appropriate age differences (52-76% of errors). Temporal shift validation maintained 99% precision, confirming model generalizability.

Conclusion

Optimized random forest models successfully identify complex family relationships from EHR demographic data with high accuracy. High-confidence predictions (≥ 0.9 probability) support automated processing for large-scale deployment, enabling population-level studies of hereditary diseases and tracking of multi-generational health outcomes.

