# Developing a Matching Algorithm for Identifying Family Units

Colby C. Uptegraft, MD, MPH[1,2,3], Alyna T. Chien, MD, MS[1], Eli Sprecher, MD, MPP[1]
[1]Boston Children's Hospital, Division of General Pediatrics, Boston, MA; [2]United States Air Force, Air Force Institute of Technology, Wright-Patterson AFB, OH; [3]Harvard Medical School, Department of Biomedical Informatics, Boston, MA

**What might the attendee be able to do after being in your session?**

This session extends the common informatics practice of patient matching to family matching. Attendees should leave with an understanding of how to build their own family matching algorithm as well as a thoughtful consideration of the many clinical and administrative use cases for leveraging this knowledge.

**Description of the Problem or Gap**

Clinical care often occurs around the silos of individual patients. While providers gather family and social histories that may be applicable for related patients, these histories are frequently inconsistent, particularly if given by different perspectives[1-3]. The care of and administration of medications or vaccines to individual patients may also depend on the health status of other members within the same household, and genetic information may inform the screening or precision medicine approach for family members[4]. With the significant influence of social factors and conditions in driving the health of individuals, knowing an individual's family unit may help providers understand some of these determinants. Additionally, healthcare administrators and social workers could use this information for care and services coordination, epidemiologists for disease and exposure tracking, and investigators for family-based research. Several major electronic health record vendors include the ability to establish family relationships; however, forming these linkages requires a manual process. This study aims to automate the identification of likely family members and operationalize the matching algorithm via a web-based application.

**Methods**

A small subset of patients (n=10,000) with current and historical data for address, phone number, and guarantor name was extracted from the Boston Children's Hospital enterprise data warehouse to build the initial family matching algorithm. This sample contained 21 individuals across 11 families. All addresses were checked and standardized via the Google Maps Platform application programming interface. Current and historical addresses, phone number, and guarantor name were matched together in a many:many fashion to create a final dataset with all possible known combinations of these data fields for each patient. This resulted in a final sample size of 19,556 for family matching. Deterministic and probabilistic matching methodologies employing different string-distance calculation techniques were trialed for individual address components, guarantor last name, guarantor first name, and phone number to maximize the matching of potential family members. The model was built with R and deployed via an RShiny web application to demo for prospective end users, which included primary care physicians, social workers, epidemiologists, research coordinators, patient access coordinators, and information technology administrators. Their feedback was consolidated and used to improve the model and web application.

**Results**

The final matching algorithm used both deterministic and probabilistic methods as this approach produced better results than either alone – 100% sensitivity and 100% specificity across the 11 families. Deterministically, records with identical phone numbers were considered a match. The algorithm then used Jaro-Winkler string-distance matching with variable weights for individual address components (street number, street, unit, & zip code), guarantor last name, guarantor first name, and phone number (1, 2, 1, 1, 5, 2.5, & 10, respectively) to probabilistically rank likely family members. Distances less than 1.5 were considered a match since this threshold achieved the highest sensitivity without loss of specificity (**Figure 1**).

**Discussion**

Through multiple iterations and diverse stakeholder engagements, this study developed a family matching algorithm with perfect accuracy. This exceptional performance is likely owed to the small sample, particularly of known families within the Boston Children's Hospital system. A much larger study with dedicated resources to identify and

validate family units will be needed to fully test and refine the algorithm before deployment in a production environment. Further discussions will also be needed around the privacy implications of linking prospective family members. However, the feedback we received for the web-based application from potential end users was overwhelmingly positive and speaks well to its future use in both clinical and non-clinical settings, assuming subsequent studies with larger sample sizes confirm its reliability.

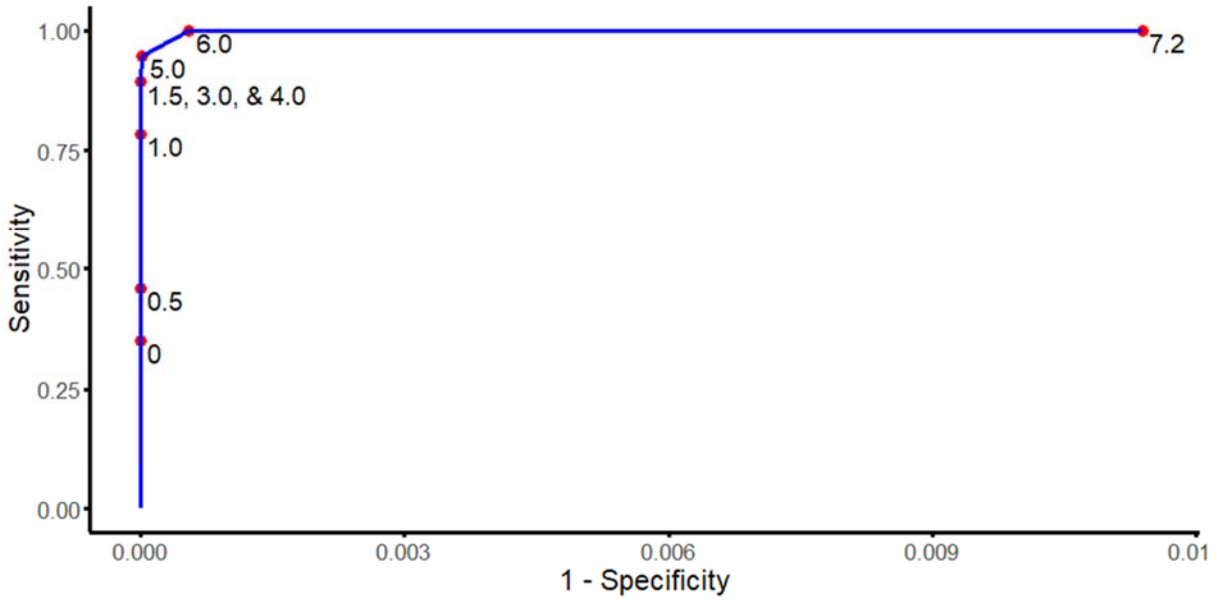


**Figure 1.** Receiving operating characteristic curve of Jaro-Winkler string distance thresholds in probabilistic component of family matching algorithm

**Conclusion**

Understanding family relationships has considerable implications for healthcare practice, research, and administration. From syncing relevant information across charts to coordinating clinical appointments and social services, access to this information offers many potential quality improvements for the provision of care and patient engagement. While larger validation studies are needed, this study presents promising results for the develop of an algorithm from readily available demographic information to automate family linkages.

**Attendee's Take-away Tool**

Attendees will leave this presentation with a detailed knowledge of how to develop a family matching algorithm at their own institution.

**Use of Knowledge Acquired at Previous AMIA Events**

I'm relatively new to attending AMIA events and have yet to encounter or attend any relevant sessions or workshops that might have contributed to this research.

**References**

1) Guttmacher AE, Collins FS, Carmona RH. The family history--more important than ever. N Engl J Med. 2004;351(22):2333–6.
2) Acheson LS, et al. Family history-taking in community family practice: implications for genetic screening. Genet Med. 2000;2(3):180–5.
3) What do parents know about the malformations afflicting the hearts of their children? Cardiol Young. 2005;15(2):125–9.
4) Frezzo TM, et al. The genetic family history as a risk assessment tool in internal medicine. Genet Med. 2003;5(2):84–91.