



The Use of Probabilistic Linkage to Remove Duplicates and Resolve Possible Matches in a Large Statewide Immunization Registry

Lawrence J Cook¹, PhD, Chris Pratt², BS, Yukiko Yoneoka², MA, Wu Xu², PhD,

1. Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, Utah; 2. Utah Department of Health, Salt Lake City, Utah



Introduction

- Immunization registries are confidential, computerized information systems that collect and consolidate vaccination data from multiple health-care providers, generate reminder and recall notifications, and assess vaccination coverage within a defined geographic area.
- Immunization registries have been used to address public health issues such as immunization rates and tracking patients during disasters.
- However, since patients have several immunizations during the first two years of life, there is a potential to falsely create multiple records for the same person.

Objective

Probabilistically identify duplicate records in a large statewide immunization database.

Methods

- Records from the Utah Statewide Immunization Information System (USIIS) were analyzed
- Vaccination records are received and processed nightly
 - A vaccination is categorized as
 - belonging to a known patient (merge)
 - belonging to a patient not currently in USIIS (insert)
 - potentially belonging to a known patient (manual review)
 - More records entering requiring manual review than existing resources can handle
 - LinkSolv 7.0.3697 was used to unduplicate the database

Potential Linkage Variables

Personal Information	
First and Middle Name	Date of Birth
SSN	Gender
Family Information	
Last Name	Mother, Father, and Guardian Names
Mother's SSN	Address / Phone # (Street, City, Zip, County)
Race / Ethnicity	Provider Information (Provider ID and Patient ID)

Missing Information	
Father's First Name	92%
County	85%
Race	85%
Mother's Maiden Name	84%
Phone Number	72%
SSN	62%
Street Address	38%

Linkage Strategies

1. Limit family information to lower risk of merging siblings
 - a. 2.1 records per address in database
 - b. Max contribution from family information limited to 0.48
2. Limit use of extreme missing variables
 - a. Use zip code to update county of residence
 - b. Reduced missing county info to less than 10% of records
3. Adjust for dependencies in linkage variables
 - a. First and middle name 40% dependent information
 - b. City, Zip, and County 60% dependent information

Results

- 1,583,706 unique patient records
 - 75,281 (5%) identified as duplicates
- 269,721 possible records
 - 96,637 (36%) probabilistic duplicates
- 173,084 unique possible patients
 - 122,191 (71%) matched to an existing patient
 - 50,893 (29%) considered new patients

Validation

- 135 records selected for expert comparison
 - Linkage failed to identify 11 (8%)
 - Due to special characters in name field
 - Identified 7 matches missed by nightly processing program

Final Linkage Variables

Personal Information	
First and Middle Name	Date of Birth
SSN (First 3 & Last 4)	Gender
Family Information	
Last Name	Mother's Maiden Name
Address (City & Zip)	County

Implications and Future Directions

- Probabilistic linkage is a feasible method for unduplicating large immunization registries.
- Results of this study have been used to inform decisions regarding the load program which received the nightly vaccination records
 - Initial results suggest greater accuracy for merging existing patients' records and fewer records requiring manual review
- Experiment will be repeated once all records have been processed by the new load program

Acknowledgements

This project was supported by the CDC funded Utah Research Center for Excellence in Public Health Informatics