University of Minnesota
Institute for Health Informatics

Creating a Large Database Test Bed with Typographical Errors for Record Linkage Evaluation

Nawanan Theera-Ampornpunt, MD, Boonchai Kijsanayotin, MD, PhD, Stuart M. Speedie, PhD
Health Informatics, University of Minnesota, Minneapolis, Minnesota

INTRODUCTION

- Health information exchange across multiple organizations requires a method or algorithm to optimally link records of the same individuals using demographic data.
- Selecting the best record linkage algorithm requires an evaluation to determine its sensitivity and specificity.
- This evaluation is facilitated by a large database test bed that closely reflects a real world population and takes into account the potential data entry errors that unfortunately occur in real-world databases.
- This study investigated the synthesis of such a database.

OBJECTIVE

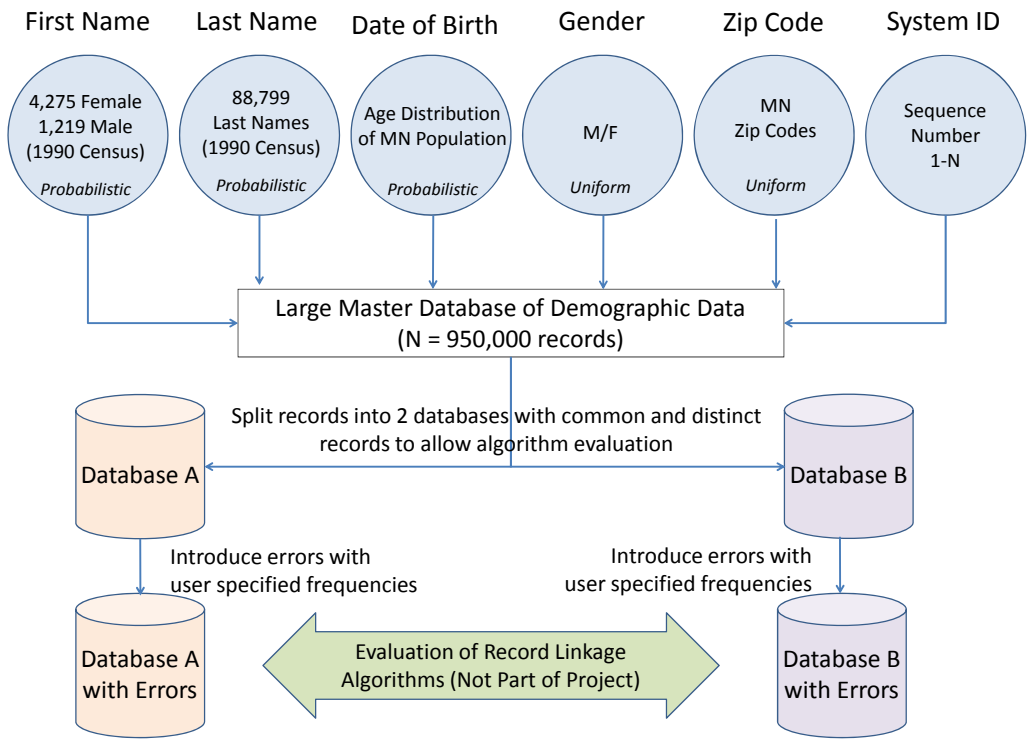
To enhance the existing methods in creating a database test bed for record linkage evaluation by developing a PHP program to:

- Create a sufficiently large database of demographic data that allows more robust and reliable evaluation of record linkage algorithms
- Model the real world distribution of key variables
- Allow users to introduce typographical errors that occur in real world due to imperfect data entry, with frequencies of error occurrence specified by the user

METHODS

Master Database Creation

- Randomly select a combination of first names and last names for each gender based on lists of names from 1990 U.S. Census publicly available and their frequencies of occurrence
- Generate date of birth based on the available age distribution of the Minnesota population and randomly select a zip code from MN zip codes using a uniform distribution



5 Types of Data Entry Errors

Character Insertion	Richard ⇨ Ricthard
Character Omission	Sullivan ⇨ Sulivan
Character Substitution	Robert ⇨ Rodert
Character Transposition	55414 ⇨ 55441
Gender Misclassification	M ⇨ F

Acknowledgment
This project was funded in part under grant number UC1 HS16155 from the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services.

METHODS (Continued)

- Generate male and female records in equal numbers, and produce a system identifier unique for each record to allow algorithm evaluation

Database Test Bed Creation

- Split records into 2 databases with both common records and distinct records

Error Introduction

- Randomly introduce errors in each applicable variable based on the frequency of each type of errors specified by the user

Next Steps: Record Linkage Algorithm Evaluation (Not Part of Study)

- Employ record linkage algorithms of interest to produce anonymous identifiers for evaluation
- Common records across the 2 databases would be used to check if an algorithm produces the same anonymous identifiers as it is supposed to.
- Distinct records across the 2 databases would be used to check if an algorithm produces different anonymous identifiers as it is supposed to.
- Errors introduced into each database can be used to assess robustness of the algorithm compared to the ideal databases with no errors.

SUMMARY OF CONCLUSIONS

- A large database test bed is achieved, allowing evaluation of record linkage algorithms
- The demographic data generated reflect real world distribution
- Data entry errors were introduced to allow algorithm evaluation of imperfect dataset