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

Abstract

Evaluation of record linkage algorithms requires a large database test bed that is representative of the real-world data. We created such a large database that reflects the demographic distribution of a typical population and contains typographical errors commonly made during data entry. This database can be used with high confidence as a test bed to evaluate various record linkage algorithms.

Introduction

Health information exchange across multiple organizations requires that there is a method or algorithm in place 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 large database.

Methods and Results

A method to synthesize a database test bed that simulates the demographic information of a real world database using Microsoft Excel was presented elsewhere. This method was further modified in order to incorporate the following features: 1) to make the database creation process scalable to potentially a million records or more, 2) to create a database that takes into account population distributions of certain demographic information, and 3) to introduce typographical errors that resemble which can occur in real-world clinical databases.

A PHP program was developed to generate a master database consisting of five demographic identifiers (first name, last name, date of birth, gender, and residential zip code). First names and last names were generated randomly based on a publicly available list of names from the 1990 U.S. Census. The list consisted of about 90% of all the valid first names and last names in the 1990 Census along with their frequencies of occurrence, with first names categorized by gender. To better reflect the real-world data, the names were selected randomly according to their frequency distribution in the Census data. In other words, names that occur frequently in the 1990 Census had a better chance of being selected. Similarly, date of birth was generated based on the available age distribution of the Minnesota population. Male and female records were generated in equal numbers. The zip code generated was randomly selected from a list of Minnesota zip codes using a uniform distribution. In addition to these five identifiers, a system identifier which was unique for each of the records was produced.

In order to evaluate how an algorithm works in a real world database that contains data entry errors, we also developed another program that introduces data entry errors into the database. The types of errors included: insertion, omission, substitution, and transposition of characters, and gender misclassification errors. The frequency of each error to be introduced can be specified by a user.

The database test bed consists of two databases with an arbitrary subset of records with known links based on the system identifier. These known links can be used as the gold standard for evaluating the sensitivity and specificity of candidate algorithms.

Conclusion

Having a large database test bed that reflects the real world in terms of demographic distributions and data entry errors provides an opportunity for researchers to evaluate record linkage algorithms with high confidence of their generalizability. Sensitivity analysis to determine how an algorithm reacts to different frequencies of errors can also be further studied.

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.

References