International Journal of Population Data Science

Journal Website: www.ijpds.org

Record Linking Techniques in the Utah Population Database to Improve Linking Rates of Hispanics

Fraser, A[1], Kinnear, S[1], and Smith, K[2]

[1]Huntsman Cancer Institute, University of Utah

[2]University of Utah

Introduction

Hispanic naming conventions frequently follow historical traditions. A person's name consists of a given name or names followed by the father's first surname and the mother's first surname or reversed if the parents wish. The challenge occurs in keying and linking these non-standard names resulting in a potential linking bias.

Objectives and Approach

Historically the Utah Population Database (UPDB) has combined multiple surnames into a single surname to standardize names such as VAN WINKLE, however this resulted in Hispanic surnames combined into nonsensical names, for example 'MARTINEZCRUZ' that were difficult to match to surnames stored in separate fields in assorted combinations. The objective of this study was to see if name specific frequencies and name arrays created with the second and third given name, maiden name and surname allowed for ultimate flexibility in matching to records which did not adhere to any standardized keying convention and resulted in better linking results.

Results

A "Gold Standard" set of Hispanic individuals with multiple record sources in UPDB and the presence of two names in the surname field were evaluated. Two Linking approaches, one using the UPDB standard methodology and the other using name arrays were compared. Both methodologies resulted in high linking rates into complete or partial sets of records per individual. Overall, the array methodology linked more records into complete sets than the standard (94.7% cf. 82.6%). Using arrays, males linked at a higher rate than females and persons from Spanish speaking countries linked at the highest rate compared with USA born or other countries. However, there was an increase in incorrect links using arrays. Name frequency distributions specific to Hispanics also proved important.

Conclusion/Implications

This study found weights based on frequencies specific to the population being linked is critical to complete linking. Using name arrays for Hispanics was most effective in males with indicators of strong ethnic ties. However, the cost of using arrays was an increase in incorrect links and further refinement is needed.

http://dx.doi.org/10.23889/ijpds.v3i4.764