International Journal of Population Data Science (2018) 3:3:089

International Journal of Population Data Science

Journal Website: www.ijpds.org

Linkage Quality Assessment for Anonymously linked Administrative Data.

Bhatt, H[1], Jagodzinski, R[1], Scott, AN[1], Twilley, L[2], and Cui, X[1]

[1]PolicyWise for Children & Families

[2]Entelechy Resources Consulting Inc.

Introduction

Linked datasets are important resources for research, but linkage errors can lead to incorrect results. For data security and privacy concerns, when linkage of personal identifiers is performed anonymously, it is difficult to assess the quality of the linked dataset. We describe the method used to perform linkage quality.

Objectives and Approach

We explored how to check the quality of linkages while preserving the privacy of individuals. We also adopted an approach that minimized time and burden on data providers involved in physical verification using randomly-generated appropriate sample sizes.

To validate these linkages, data providers were given random samples of 50 unique records from both linked and unlinked individuals across two other Government programs. Data providers were asked to look at the records associated with those individuals in their original datasets. Three types of linkage results were validated: cross-program linkages, cross-program non-linkages, and within-program linkages. Proportions of false-matches and missed-matches were estimated.

Results

Twenty data providers checked their samples with two other programs which gave us a sample of 2000 individuals. The linkage process, based on anonymized personal identifiers, resulted in high true positive and high true negative rates. Agreement between human judges and the linkage software was strong. Results of this exercise and other linkage validation examinations provided confidence in the accuracy of the linkage process. With false matches occurring approximately only 3% of the time and virtually no missed-matches occurring, no adjustments were deemed necessary. Although linkage rates were reassuring, the sample sizes used for comparison were small, so it is expected that there would be significant variation associated with this 3% estimate; caution is advised in its use.

Conclusion/Implications

Proportions of false-matches and missed-matches determine linkage quality which is the base for research when linkages are performed anonymously. A low proportion of false-matches and an absence of missed-matches was an indication of robust linkages.

http://dx.doi.org/10.23889/ijpds.v3i4.679